BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

As rescanning documents *will not* correct images, please do not report the images to the Image Problem Mailbox.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 January 2002 (17.01.2002)

PCT

(10) International Publication Number
WO 02/04629 A2

(51) International Patent Classification[7]: C12N 15/10

(21) International Application Number: PCT/US01/21532

(22) International Filing Date: 5 July 2001 (05.07.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/216,798 7 July 2000 (07.07.2000) US

(71) Applicant *(for all designated States except US)*: **MAXYGEN, INC.** [US/US]; 515 Galveston Drive, Redwood City, CA 94063 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **DELCARDAYRE, Stephen** [US/US]; 2049 Monroe Avenue, Belmont, CA 94002 (US). **PATNAIK, Ranjan** [IN/US]; 5273 Mill Creek Lane, San Jose, CA 95136 (US). **PATTEN, Phillip** [US/US]; 261 La Cuesta Drive, Menlo Park, CA 94028 (US). **TOBIN, Matthew** [US/US]; 5662 Sunflower Lane, San Jose, CA 95118 (US). **NESS Jon, E.** [US/US]; 1220 N. Fairoaks Avenue, #2115, Sunnyvale, CA 94089 (US). **COX, Anthony** [GB/US]; 1730 Plaza Court, Mountain View, CA 94040 (US). **GIVER, Lorraine, J.** [US/US]; 2538 Hawkington Court, Santa Clara, CA 95051 (US). **MCBRIDE, Kevin** [US/US]; 1309 Marina Circle, Davis, CA 95616 (US). **ZAHN, Kenneth** [US/US]; 707 Leahy Street, #315B, Redwood City, CA 94061 (US).

(74) Agents: **QUINE, Jonathan, Alan** et al.; The Law Offices of Jonathan Alan Quine, P.O. Box 458, Alameda, CA 94501 (US).

(81) Designated States *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) Designated States *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published:
— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 02/04629 A2

(54) Title: MOLECULAR BREEDING OF TRANSPOSABLE ELEMENTS

(57) Abstract: Methods for producing transposable elements with improved properties as vectors are provided. Directed evolution procedures are employed to improve characteristics of transposable elements, including transposons and insertion sequences as vectors. Methods for generating diversity in vivo and in vitro using transposable elements as vectors are provided.

MOLECULAR BREEDING OF TRANSPOSABLE ELEMENTS

CROSS REFERENCE TO RELATED APPLICATIONS

This application claims priority to and benefit of United States Provisional Application Number 60/216,798, filed July 7, 2000, the specification of which is incorporated herein in its entirety for all purposes.

BACKGROUND OF THE INVENTION

Industrial production of many biochemicals is currently achieved through use of whole cells as biocatalysts or by fermentation. Economic production of these chemicals are typically dependent on the productivity of the biocatalyst under process conditions, which generally tend to be significantly different than the conditions for which the biocatalyst has naturally evolved. The current technology used to engineer strains to be more productive under desired process conditions generally involves one or both of: various forms of mutagenesis on a host organism coupled with screens and selections and/or overexpression of desired enzymes using standard molecular biology tools.

Although the above methods are successful to a certain extent, many limitations and disadvantages exist. For example, classical mutagenesis and screening procedures are time consuming, and in most cases, improvements observed in one host cannot be transferred to another host due to lack of significant knowledge about the relevant genetic interactions in the host and recipient species. In cases where genetic methodology is used, only pair-wise recombination of useful mutations can be assessed at any one time. Briefly, the synergistic effect of many useful mutations on a desired phenotype cannot be assessed conveniently using current methods due to the difficulty in assessing the mutations in combinatorial fashion.

Typically, in a classical strain improvement program, many desirable phenotypes are observed in different host backgrounds but the ability to combine these phenotypes into a single production strain is severely limited due to lack of methodology for inter-species genetic exchange, low homologous recombination efficiency, low

electroporation efficiency in certain cases and most importantly lack of a suitable method for creating combinatorial genomes.

The evolution of microbial genomes is catalyzed by the processes of horizontal gene transfer. Indeed, these processes most closely resemble the exchange of genetic information that occurs during the sexual cycle of eukaryotic organisms. Natural competence, general transduction, conjugation, and transposon mediated gene exchange all contribute to horizontal gene transfer. Insertion sequences and transposons are found distributed throughout most genomes thus far investigated. The mobilization of IS elements and transposons within and between genomes is a primary mechanism for the reorganization of genome structure and the horizontal exchange of genetic information.

The goal of rapidly evolving whole microbial cells by "whole genome shuffling" will most efficiently be realized when the natural mechanisms by which microbial cells evolve can be harnessed and accelerated in a laboratory setting. Described here is a general approach to microbial breeding that exploits the efficiency of transposons to mobilize and insert large pieces of heterologous DNA into the chromosome of a broad range of microbial hosts. This mechanism of genetic exchange employs non-homologous recombination and provides a means by which divergent heterologous DNA can be incorporated into the genome of an unrelated host. Extensive processes for whole genome shuffling are found in USSN 09/116,188 "Evolution of Whole Cells and Organisms by Recursive Recombination" by del Cardayre et al. filed July 15, 1998 and PCT publications WO 00/04190 "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination," by del Cardayre et al. published 1/27/2000. The present invention provides additional improvements in horizontal gene transfer vectors and artificial evolution methods.

SUMMARY OF THE INVENTION

The present invention provides methods for producing transposable elements, including transposons and insertion sequences, with improved properties. In general, the methods of the invention involve diversifying, e.g., recombining, polynucleotide segments corresponding to one or more component of a transposable element to produce a library of recombinant transposable element components. The library is then evaluated to identify members with improved properties. Optionally, the

process is performed in a recursive fashion. In some embodiments, the transposable element is recovered following transposition into the host cell.

For example, substrates for diversification, e.g., recombination, or "shuffling" reactions can include any component of a transposable element, such as a transposase or an inverted repeat. Alternatively, only a subsequence of such a component provides the basis for recombination. In other cases, multiple components, including entire transposable elements, e.g., mini-transposons, mini-IS elements, etc., are recombined, e.g., shuffled simultaneously. Suitable substrates for the methods of the present invention include transposable elements derived from a variety of sources, including bacterial, fungal, plant and animal transposable elements. Such transposable elements can be broadly categorized based on their mechanism of transposition into Class I, e.g., retrotransposons, retroposons, and SINE-like elements, e.g., *Ty-1*, *Copia*, *gypsy*, and the like, and Class II, e.g., *Fotl/Pogo*, *Tcl/Mariner*, etc. Both Class I and Class II transposable elements are substrates of the invention. In certain preferred embodiments, transposable elements that are TN3, TN5, TN10, TN917, ISS1, TN5990, Ty1, Ty2, Ty3, and *mariner* are substrates for the diversification, e.g., shuffling methods of the invention. Diversification, e.g., shuffling of the transposable element sequences is performed in vitro, in vivo, in silico, or any combination thereof.

The methods of the present invention are used to produce transposable elements with a variety of improved properties; in particular, with respect to their performance as delivery vectors. Desirable properties include: altered specificity of integration, host adaptation, increased or decreased recombinase activity, increased or decreased transposase activity, increased or decreased recombinase specificity, increased or decreased transposase specificity, increased or decreased size of exogenous DNA transposed, increased or decreased copy number, increased or decreased efficiency of transposition, increased or decreased preference for episomal targeting, increased or decreased preference for chromosomal targeting, increased efficiency of integration into non-supercoiled DNA, and increased efficiency of in vitro transposition.

In general, transposable elements, or their components with desired properties are identified by one or more selection or screening protocols. In one preferred embodiment, components of transposable elements that mediate in vitro transposition with increased efficiency are identified by evaluating in vitro transposition reactions comprising a transposase, a donor polynucleotide having an inverted repeat, and a target

polynucleotide, of which one or more components results from diversification procedures, e.g., shuffling. In some embodiments, the in vitro transposition reactions include transposomes.

In another preferred embodiment, transposable elements that transpose with increased efficiency in a specified host cell type are identified by introducing a plurality of transposable elements, differing by at least one nucleotide, into a population of host cells, and selecting host cells that have integrated the transposable element into a chromosome or episome. Such methods are facilitated by the use of a transposable element including, in the direction of transcription: (a) a polynucleotide comprising a transcription regulatory sequence; (b) a 5' splice donor site; (c) a first inverted repeat; (d) a 3' splice acceptor site; (e) a polynucleotide encoding a transposase; (f) a polynucleotide encoding a selectable marker; and (g) a second inverted repeat. In some embodiments the transposase is transiently expressed preceding transposition. Following transposition, e.g., integration, host cells expressing a sufficient level of a marker, e.g., antibiotic resistance, encoded by the transposable element are selected. In certain embodiments, the selected host cells are mammalian cells. In some cases, the transposable element is a *Mariner*-like transposable element, having a *Mariner* transposase and *Mariner* inverted repeats.

In some embodiments, sequences comprising a transposable element are incorporated into a recombinant vector such as a recombinant episomal vector, e.g., a plasmid. In one embodiment, the vector is a delivery vector. The delivery vector has an origin of replication active in one or more cloning hosts, as well as a conditional origin of replication active in a selected target cell; at least one screenable or selectable marker, e.g., antibiotic resistance, toxicity resistance, conferred prototrophy; and a mini-transposon having inverted repeats flanking a multicloning site (MCS) and a transposase operably linked to a promoter active in the selected target cell. In certain preferred embodiments, the transposase is derived by a directed evolution process. In some embodiments, the sequences encoding the transposase are situated in close proximity to an end of the mini-transposon.

Such recombinant delivery vectors are also an aspect of the invention. Exemplary replication origins of the vectors include origins derived from: ColE1, pACYC, p15A, RK4, RK6, pCM595, pSa, pUB110, pE194, pG+, 2 micron circles, and artificial chromosomes. Temperature sensitive origins of replication favorable in the

vectors of the present invention include pSA3, pE194, and pG+tm. Mini-transposons derived from transposons or insertion sequence elements including insertion sequences and their components including inverted repeats and transposases selected from among: IS1, IS2, IS3, IS4, IS5, IS6, IS10, IS21, IS30, IS50, IS91, IS150, IS161, IS186, IS200, IS903, IS3411, IssHO1, IS600, IS22, IS52, IS222, IS401, IS402, IS403, IS404, IS405, IS411, IS476, IS60, IS66, IS426, IS492, IS4400, ISR1, ISRm1, ISRm2, RSRj-alpha, RSRj-beta, IS701, IS 231, IS2150, IS256, IS431, IS257, ISS1, IS110, IS466, ISL1, and Gamma delta, are all favorably employed in the context of the present invention.

Similarly, transposons from a variety of sources including conjugative transposons, e.g., Tn916, Tn918, Tn919, Tn925, Tn1545, 3951, and BM6001 element; Class II transposons, e.g., TN551, Tn917, Tn3871, Tn4430, Tn4556, Tn4451, Tn4452; and other transposons, e.g., Tn554, Tn3853; Tn4001, Tn3851, Tn552, Tn4002, Tn3852, Tn4201, and Tn4003 TN3, TN5, TN10, TN917, ISS1, TN5990, Ty1, Ty2, Ty3, and *mariner* are favorably employed as mini-transposons in the recombinant delivery vectors of the invention.

Transposable elements with improved characteristics are a feature of the present invention. Similarly, components, e.g., transpsosases, integrases, inverted repeats, etc., of transposable elements conferring improved characteristics are a feature of the invention. Transposable elements having (and transposable element components conferring) such desirable properties as altered specificity of integration, host adaptation, increased or decreased recombinase activity, increased or decreased transposase activity, increased or decreased recombinase specificity, increased or decreased transposase specificity, increased or decreased size of exogenous DNA transposed, increased or decreased copy number, increased or decreased efficiency of transposition, increased or decreased preference for episomal targeting, increased or decreased preference for chromosomal targeting, increased efficiency of integration into non-supercoiled DNA, and increased efficiency of in vitro transposition are produced by the methods of the invention.

In another aspect, the invention provides methods for producing a transposase that efficiently catalyzes in vitro tranposition. A population of polynucleotide segments encoding one or more transposases or subportions of one or more transposase are recombined to produce a library of variant transposases. The variant transposases are then evaluated for their ability to efficiently catalyze in vitro transposition. In an

embodiment, variant transposases that efficiently catalyze in vitro transposition are identified by incubating a plurality of in vitro transposition reactions under conditions permissive for in vitro transposition, and identifying those reactions that proceed with greater efficiency than an in vitro transposition reaction mediated by a parental transposase. In vitro transposition reactions include: a variant transposase encoded by a member of the library of recombinant polynucleotides; a donor polynucleotide with at least one inverted repeat (e.g., one, two or a number sufficient for transposition); and a target polynucleotide. Transposases produced according to the methods are also a feature of the invention. In preferred embodiments, the transposases are derived by a directed evolution process from transposases of one or more of TN3, TN5, TN10, TN917, TN5990, ISS1, Ty1, Ty2, Ty3 and mariner. Similarly, reaction mixes and cells including the transposases produced by the methods of the invention are an aspect of the invention.

Another aspect of the invention relates to the generation of diversity in a population of nucleic acids. The invention provides methods of generating diversity in a population of nucleic acids by contacting a recombinant, e.g., shuffled transposable element, or a shuffled component of a transposable element with a plurality of subject nucleic acids under conditions permissive for transposition. Alternative embodiments involve contacting the transposable element, or transposable element component, and the subject nucleic acids in vitro or in vivo. In one embodiment, altered subject nucleic acids are identified.

In some embodiments, the recombinant, e.g., shuffled transposable element component is a transposase. In an embodiment, a transposome made up of a recombinant, e.g., shuffled transposase bound to a donor nucleic acid having sequences recognized by the shuffled transposase is introduced into a cell, e.g., by electroporation. In alternative embodiments, the transposome is contacted with the subject nucleic acids in an acellular reaction mix.

In another aspect, the invention provides methods for generating diversity in a population of nucleic acids in vitro using transposomes. Transposomes incorporating a diverse (e.g., from multiple species or strains of microorganism) library of donor nucleic acids having transposase recognition sites are recombined in vitro with a population of acceptor nucleic acids. Optionally, the recombinant nucleic acids are introduced into cells and cells expressing a desired phenotype is screened or selected. In

some embodiments, the recombination process is performed recursively, with or without intervening screening or selection steps.

The invention further provides methods for identifying chromosomal loci that generate a desired level of gene expression. Generally, such methods involve (i) transfecting a plurality of host cells expressing a transposase with a vector characterized by inverted repeats flanking a promoter, a site specific recombinase recognition site, and one or more screenable or selectable marker; (ii) selecting host cells that have integrated the vector and express a sufficient level of a selectable marker encoded by the vector to survive selection; and (iii) evaluating the surviving host cells for a desired level of expression of a marker. Such vectors are a feature of the invention. For example, in the case of identifying a locus in a chromosome of a selected mammalian cell line expressing, e.g., a *Mariner* transposase, the inverted repeats of the vector are preferably derived from a transposable element, e.g., *Mariner*, the site specific recombinase recognition site comprises a loxP site, and the promoter comprises, e.g., a cytomegalovirus (CMV) promoter active in the selected cell line.

In preferred embodiments, the transposase is a recombinant, e.g., shuffled transposase with at least one improved property, e.g., sequence specificity, activity level, species selectivity, allostery, control, etc., relative to a parental transposase from which it is derived. In some embodiments, the vector also supplies expression of the transposase by including a polynucleotide encoding the transposase operably linked to a promoter functional in the host cells. Alternatively, the transposase activity is supplied by an additional vector, or integrated into a chromosome. In some embodiments, the transposase is transiently, e.g., inducibly, expressed. In some cases, a polynucleotide of interest is integrated into the chromosomal locus previously identified and integrants are identified exhibiting a desired level of expression of the gene of interest.

The present invention also provides, e.g., a transposable element comprising, in the order of transcription: an int encoding sequence and an xis encoding sequence, each operably linked to a promoter functional in the target cell; a mini-IS element; an origin of replication functional in a cloning host, a first and a second selectable marker; and a second, temperature sensitive, origin of replication functional in the target cell, is a feature of the invention.

BRIEF DESCRIPTION OF THE DRAWING

Figures 1A-1C are schematic illustrations of recombinant vectors incorporating transposable elements.

Figures 2A-2B are schematic illustrations of transposon vectors.

Figure 3 is a schematic illustration of a continuous fermentation protocol for selecting variants with a desired phenotype.

Figures 4A-4D schematically illustrate in vitro transposome mediated recombination.

DETAILED DISCUSSION OF THE INVENTION

The present invention relates to the production of transposable elements with improved characteristics, most particularly, with respect to their function as vectors for genetic manipulation. Nucleic acid diversification procedures, such as shuffling are used to recombine and/or mutate naturally occuring, mutant and/or artificial polynucleotides corresponding to transposable elements and their components, e.g., repeat sequences, transposases, regulatory sequences and the like. Following generation of a library of recombinant transposable element sequences, transposable elements and transposable element components that exhibit desired properties are identified through a variety of screening and selection procedures. Transposable elements with novel and enhanced properties are valuable as vectors for delivering DNA into cells, and for generating diversity within a population of cells by transposition mediated events. In addition, isolated components, e.g., transposases are valuable as tools for mediating DNA delivery and recombination both in vitro and in vivo.

DEFINITIONS

Unless defined otherwise, all scientific and technical terms are understood to have the same meaning as commonly used in the art to which they pertain. For the purpose of the present invention the following terms are defined below.

A "transposable element" (TE) or "transposable genetic element" is a DNA sequence that can move from one location to another in a cell. Movement of a transposable element can occur from episome to episome, from episome to chromosome, from chromosome to chromosome, or from chromosome to episome. Transposable elements are characterized by the presence of inverted repeat sequences at their termini. Mobilization is mediated enzymatically by a "transposase."

Structurally, a transposable element is categorized as a "transposon," ("TN") or an "insertion sequence element," (IS element) based on the presence or absence, respectively, of genetic sequences in addition to those necessary for mobilization of the element. A mini-transposon or mini-IS element lacks sequences encoding a transposase.

In the context of the present invention, a "component" of a transposable element refers to any identifiable functional unit, e.g., polynucleotide repeats, transposase, whether nucleic acid or protein, of a transposable element. A "subportion" of a transposable element or transposable element component refers to any subsequence of a transposable element or transposable element homolog, including artificial sequences, up to and including an entire transposable element or transposable element component.

A "parental" transposable element or transposable element component, e.g., transposase, refers to a transposable element, or component, that is provided as a substrate for a directed evolution process, e.g., nucleic acid shuffling, according to any of the formats described herein. Typically, such a substrate is provided in actual (e.g., in vitro, in vivo shuffling) or virtual (e.g., in silico shuffling) form as a polynucleotide "segment."

An "in vitro transposition reaction" is a recombination between nucleic acid substrates, e.g., a donor DNA molecule and a target DNA molecule, mediated by a transposase in an acellular reaction mixture. The term "transposome," or "synaptic complex," refers to a functional complex made up of a transposase associated with a transposable polynucleotide via specific recognition sequences, e.g., inverted repeat sequences.

"Screening" is, in general, a two-step process in which one first determines which cells, organisms or molecules, do and do not express a detectable marker, or phenotype (or a selected level of marker or phenotype), and then physically separates the cells, organisms or molecules, having the desired property. "Selection" is a form of screening in which identification and physical separation are achieved simultaneously by expression of a selectable marker, which under some circumstances, allows cells expressing the marker to survive while other cells die (or vice versa). Screening reporters include visible markers such as luciferase, β-glucuronidase, green fluorescent protein (GFP) as well as functional attributes evaluated according to a variety of specific assays. Selectable markers include antibiotic and herbicide resistance genes. A special class of

selectable markers are negatively selectable markers. Cells or organisms expressing a negatively selectable marker die under appropriate selection conditions while organisms lacking or having a non-functional form of the marker survive.

The present invention provides methods, characterized as artificial or directed evolution, for evolving transposable elements and components thereof to acquire desired properties. Directed evolution involves the generation of sequence diversity in a nucleic acid, or population of nucleic acids, followed by or interspersed with screening or selection procedures to identify nucleic acids with desired structural or functional properties or characteristics. The invention utilizes, e.g., MolecularBreeding™ technologies, in a process of directed evolution, to generate and optimize mutations resulting in transposable elements with improved characteristics, e.g., as vectors and mutagenic agents. The resultant transposable elements and components, e.g., transposases, are used to introduce and/or mobilize polynucleotides into or within a genome in a wide variety of applications.

In a general format, polynucleotide segments corresponding to a transposable element or a component of a transposable element, or to a subportion thereof, are recombined, in vitro, in vivo, or in silico to produce a library of recombinant transposable element polynucleotides. The polynucleotide segments provided can be physical, such as isolated DNAs derived from naturally occurring transposable elements or synthesized oligonucleotides corresponding to (or complementary to) a portion of a wild type or variant transposable element or component thereof. Alternatively, the polynucleotide segments can be virtual, e.g., in silico representations of a naturally occurring or synthetic DNA sequence stored in a computer readable medium.

The polynucleotide segments are recombined, and optionally mutated, one or more times to generate a library of recombinant transposable element polynucleotides. The recombination process can be performed in vitro, in vivo, or in silico, or in any combination of formats as described in further detail herein and in the cited references. The library is then evaluated, by a variety of techniques available in the art chosen to identify recombinants with the desired property.

For example, polynucleotide segments that are fragments derived by DNAse digestion from a transposable element isolated from a given bacterial or eukaryotic species can be combined in vitro with synthesized degenerate oligonucleotides corresponding to a variety of naturally occuring or artificial sequences, some or all or

none of which correspond to sequences of known transposable elements. The segments are then recombined according to any of the procedures described herein, or in the cited references. For example, the DNAse generated segments described above can be recombined based on homology by PCR reassembly protocols previously described by the inventors and their coworkers.

Alternatively, in silico character strings representing polynucleotides of any number of transposable element and other sequences, e.g., recombinases, integrases, etc., can be recombined by a computer according to genetic algorithms that do not rely on homology. Optionally, the resulting recombinant polynucleotides can be synthesized, and if desired, subject to additional rounds of recombination in vitro or in vivo.

In some cases, the polynucleotide segments are recombined in the context of a recombinant vector. In other cases, individual components or transposable elements are recombined and subsequently recovered, e.g., by a polymerase chain reaction (PCR), ligase chain reaction (LCR), Qβ-replicase amplification, NASBA or cloning. Upon recovery, it is often desirable to conserve and/or reproduce the component or transposable element in the context of a vector.

Transposable elements, transposable element components and vectors comprising transposable elements, produced by the methods of the invention, are used to alter the genomes of cells and organisms both as mutagenic agents and as recombinant delivery vectors. In the former case, transposable elements with improved characteristics as mutagens, e.g., increased transposase activity, increased recombinase activity, decreased transposase specificity, decreased recombinase specificity, increased copy number, increased efficiency of transposition, etc., are introduced into cells where they are constitutively or inducibly activated to undergo transposition events. This provides the basis for novel and improved methods for generating diversity both in vitro and in vivo. In the latter case, transposable elements of the invention that are delivery vectors are employed to introduce sequences of interest into the genome of a cell (or organism). In addition, these methods are useful for the creation of combinatorial genomes.

Additionally, specialized vectors that include transposable elements and transposable element components useful for genetic manipulation are described. For example, vectors and methods useful for identifying a chromosomal locus capable of supporting a desired level of gene expression are provided, as are methods for integrating a gene of interest into such a locus.

TRANSPOSABLE ELEMENTS

Transposable elements are DNA sequences that can move between locations within a genome, and in some cased between genomes. Transposable genetic elements have been identified in a wide range of organisms, including both prokaryotes and eukaryotes, and since their identification have found numerous uses as vectors, markers, and as mutagens. Transposable elements, as a group, share certain advantageous features that make them particularly well suited as agents of genetic change.

In general, transposable elements that include only sequences necessary for transposition are designated "insertion sequence (IS) elements," or "insertion sequences." IS elements contain genes encoding proteins necessary for transposition, (i.e., excision and insertion) flanked by short inverted repeats. In contrast, a "transposon" (TN) typically incorporates genetic sequences in addition to those involved in mobilizing the DNA. Often these additional sequences confer resistance to antibiotics or produce toxins. The conversion of an IS element to a transposon can occur when two IS elements surrounding a region of genomic DNA excise together mobilizing the intervening genomic DNA. Conjugal transposons further encode the ability to catalyze the conjugal transfer of the excised transposon to a different cell where it integrates into the chromosome.

Both IS elements and transposons are the subject of the present invention. IS elements can be readily adapted, e.g., as vectors for DNA delivery, through the introduction of a multiple-cloning site (MCS). Similarly, DNA sequences, e.g., genes of interest, can be engineered into transposons either as replacements for, or in addition to, sequences non-essential for mobilizing the transposon. Regardless of whether an IS element or transposon is selected, the transposable element can be manipulated according to the methods described herein to acquire novel and desirable properties.

Transposable elements can be categorized into two broad classes based on their mode of transposition. These are designated Class I and Class II; both have applications as mutagens and as delivery vectors, and both are subject to improvement by the methods of the invention. Class I transposable elements transpose by an RNA intermediate and use reverse transcriptases, i.e., they are retroelements. There are at least three types of Class I transposable elements.

Retrotransposons of the *Ty-1/Copia* family and the *gypsy* family. Retrotransposons typically contain LTRs, and genes encoding viral coat proteins (gag) and reverse transcriptase, RnaseH, integrase and polymerase (pol) genes.

Retroposons (LINE-like retroelements) have poly-A tails but do not have LTRs, and intact retroposons also contain *gag* and *pol*.

SINE-like elements are derived from transcripts of RNA polymerase III. They do not contain *gag* or *pol* or LTRs, and are trans-activated by RTs from the retroelements or retrotransposons.

Class II transposable elements transpose directly at the DNA level, and include the *Fot1/Pogo* or *Tc1/Mariner* families, among others. Class II transposons have short inverted repeats and often encode transposases of different types.

Transposition occurs by either a conservative or replicative mechanism depending on the transposable element.

So-called "Mini-transposons" lack transposases altogether, and can be constructed to permit provision of the transposase in trans.

Transposable elements are distributed throughout the genomes of a wide variety of species, including both prokaryotes and eukaryotes. Depending on the application, and in particular on the host cell to be the subject of manipulation by the transposable elements of the invention, a choice is made from among the myriad transposable elements.

Bacterial Transposable elements

Bacterial cells are especially amenable to genome manipulation, e.g., diversification, using transposable elements. Transposons and insertion sequences have been isolated and characterized from numerous gram-negative and gram-positive bacterial species, and bacterial TEs of both Class I and Class II varieties, and that are conjugative transposons are favorably employed in the methods of the invention. Of these, both insertion sequence elements and transposons have been cloned and characterized. Insertion sequences are typically between about 0.7 and 2 kb, while transposons range in size to greater than 50 kb. A number of references provide extensive lists of sources of sequences suitable in the context of the present invention (*see*, e.g., Galas and Chandler, *Bacterial Insertion Sequences*; Murphy, *Transposable elements in gram-positive bacteria*). The following are provided by way of illustration and not by limitation, as it will be readily understood that sequences derived or inferred from any

transposable element, whether naturally occurring, mutant or artificial, can be recombined according to the methods of the invention to produce transposable elements with desired characteristics.

For example, insertion sequences and their components including inverted repeats and transposases selected from among: IS1, IS2, IS3, IS4, IS5, IS6, IS10, IS21, IS30, IS50, IS91, IS150, IS161, IS186, IS200, IS903, IS3411, IssHO1, IS600, IS22, IS52, IS222, IS401, IS402, IS403, IS404, IS405, IS411, IS476, IS60, IS66, IS426, IS492, IS4400, ISR1, ISRm1, ISRm2, RSRj-alpha, RSRj-beta, IS701, IS 231, IS2150, IS256, IS431, IS257, ISS1, IS110, IS466, ISL1, and Gamma delta, are all favorably employed in the context of the present invention.

Similarly, transposons from a variety of sources including conjugative transposons, e.g., Tn916, Tn918, Tn919, Tn925, Tn1545, 3951, and BM6001 element; Class II transposons, e.g., TN551, Tn917, Tn3871, Tn4430, Tn4556, Tn4451, Tn4452; and other transposons, e.g., Tn554, Tn3853; Tn4001, Tn3851, Tn552, Tn4002, Tn3852, Tn4201, and Tn4003 are all favorable in the context of the present invention.

Fungal Transposable elements

The full range of known eukaryotic transposable elements is observed in fungal genomes, including Class I and Class II transposons (for recent reviews, *see*, e.g., Kempken and Kuck (1998) Bioessays 20:652; Daboussi (1997) Genetica 100:253; US Patent No. 5,985,570 "Identification of and Cloning a Mobile Transposon from *ASPERGILLUS*" to Amutan et al., issued Nov. 16, 1999). Evidence of transposons is frequently observed in pathogenic species, and "untamed" species in general. Multiple copies of transposons frequently exist in a fungal genome, resulting in genetic instability (sometimes referred to as "genomic plasticity") due at least in part to stimulation of genome reorganization by transposon activity.

Filamentous fungi are unusual in that they often contain multiple nuclei per cytoplasmic compartment (are coenocytic). Cells containing genetically different nuclei are designated heterkaryons, and are formed via anastamosis (fusion of hyphae). Transpositons that would lead to lethality or other detrimental effects in a mononuclear cell are often capable of surviving in a heterokaryotic cell. This provides the significant benefits of retaining mutations that would otherwise be lost, and permitting the involvement of such mutations in genome evolution. For example, the Tad LINE-like element (of *N. crassa* has been shown to transpose through a cytoplasmic intermediate

between heterokaryon nuclei, and can introduce itself rapidly into new genomes. This is particularly useful in the application of a pool-wise recombination format.

Some fungal species can inactivate incoming transposons, e.g., through processes designated "RIP" (repeat induced point mutagenesis) and "MIP" (methylation induced premeiotically). In *Neurospora crassa* RIP causes C-to-T transitions in repeat sequences at a high frequency (*see,* e.g., Selker (1998) Proc Nat'l Acad Sci USA 95:9430; and references therein). MIP causes methylation of cytosine in DNA repeats in *Ascobolis immerses* (Rossignol and Faugeron (1994) Experientia 50: 307). Most fungal species having transposons lack an obvious sexual cycle (or, have one that is only rarely active). In these cases RIP and MIP is not generally a problem as long as a cross is not achieved.

The following list of exemplary fungal TEs includes elements with a Class I transposition mechanism, e.g., Hideaway, MARS1, MARS2, MARS3, MARS4, MARS5, Afut1, Boty, Cft-1, CfT1, EGH24-1, Eg-R1, Foret-1, Palm, Skippy, Repa, Fosbury, Grasshopper, Maggy, MGR583, Mg-SINE, MGSR1, Nrs1, Pogo, Tad1-1; and transposons with a Class II transposition mechanism including, Ascot-1, Tascot, F2P08, Ant1, Tan, Vader, Restless-d1, Flipper, Fcc1, Fot1, Fot2, Impala, Hop, MGR586, Pot3, Pot2, Nht1, Guest, Pce1, PSR, and Restless.

Transposable elements have likewise been isolated from yeast (Saccharomyces cerevisiae) and are favorable in the context of the present invention. Such elements include Ty1, Ty2, Ty3, as well as δ, σ, τ, and Ω elements.

Transposable elements in other eukaryotes

In addition to the previously enumerated transposable elements, numerous transposable elements have been characterized from multicellular eukaryotes, including both plants and animals. For example, numerous retrotransposons have been described in plant species. Such retrotransposons mobilize and translocate via a RNA intermediate in a reaction catalyzed by reverse transcriptase and RNase H encoded by the transposon. Examples fall into the Ty1-*copia* and Ty3-*gypsy* groups as well as into the SINE-like and LINE-like classifications. A more detailed discussion can be found in Kumar and Bennetzen (1999) *Plant Retrotransposons* in Annual Review of Genetics 33:479. In addition DNA transposable elements such as Ac, Tam1 and En/Spm are also found in a wide variety of plant species, and can be utilized in the present invention.

Similarly, many transposons useful in the context of the present invention have been identified in animal species. To date, active transposons have been isolated

from invertebrate species, while inactive elements have been found in several vertebrate genomes. For a recent review, *see*, Plasterk and Izsvak (1999) *Resident aliens* in Trends in Genetics 15:326. In particular, transposons of the Tc1/*mariner* and *Fot/Pogo* groups can be favorably utilized in the present invention. For example, various inactive elements, from a single host species, or from several species, any number of which can be active or inactive in their respective hosts, can be recombined according to any of the recombination formats described herein, and selected for a desirable level of transposition activity in a target cell type.

EVOLVING TRANSPOSABLE ELEMENTS WITH DESIRED PROPERTIES

Sequences derived from any of the above, or other, transposable elements can be recombined and the recombinant products evaluated for the acquisition of desired properties. Among the many properties that can be achieved by the methods of the invention are increased or decreased specificity of integration, host adaptation, increased or decreased recombinase activity, increased or decreased transposase activity, increased or decreased recombinase specificity, increased or decreased transposase specificity, desired size of the exogenous DNA transposed, copy number of integrated elements, increased or decreased efficiency of transposition, increased or decreased preference for episomal targeting, increased or decreased preference for chromosomal targeting, increased efficiency of integration into non-supercoiled DNA, and increased efficiency of in vitro transposition, etc. Numerous assays useful for detecting transposable elements and their components with these and other properties are available to one of skill in the art.

In many cases, desired outcomes can be achieved by focusing the recombination process on an individual component of the transposable element. The following series of illustrative examples demonstrates how individual components of transposable elements can be evolved to acquire a subset of pre-determined characteristics. These examples are provided to facilitate and not to limit the present invention. In general, the identification of recombinant polynucleotides with the specified qualities is dependent on the selection or screening protocol employed. Thus, a number of different desired properties can be selected or screened simultaneously from among the same library of recombinant polynucleotides. Indeed, such simultaneous evaluation for multiple properties can be advantageously employed to identify

recombinant polynucleotides that are improved with respect to multiple properties when compared to the parental sequences that were the subject of the diversification reactions.

Specificity of integration site

The inverted repeats flanking an IS element or transposon are recognized by the transposable element's transposase and influence the sequences into which the element will transpose. Some ISs and TNs are very specific for a particular target sequence and thus integrate into a genome relatively non-randomly, i.e., with site specificity. Others are less specific and integrate in an essentially random manner. The Inverted repeats (e.g., derived from a variety of naturally occuring or mutant transposable elements, or artificially synthesized degenerate oligonucleotides) of ISs and TNs can be recombined, e.g., shuffled, mutated or otherwise modified and screened for a change in specificity, i.e., either more specific integration or more random integration. These sequences can also be shuffled, mutated or diversified by other diversity generating method, and screened for the ability of a new IS or TN incorporating the diversified repeats to efficiently transpose in a new host. For example, a library of TNs differing in the sequences of their inverted repeats are delivered to a target cell or organism of choice. To screen for an increase in the specificity of integration, a screening method involving the detection of integration into a pre-determined sequence can be used. For example, a specific target sequence, such as green fluorescent protein (GFP), is introduced into a chromosome or episome maintained in the chosen cell. Cells losing fluorescence are enriched for those having TN integrations into the target sequence within the GFP gene. TNs having integrated into the target sequence are selectively amplified from a pool of the gDNA isolated from the non fluorescent colonies by PCR. The primers used in this reaction are hybrid sequences of the inverted repeats and the target sequence. In this manner, only TNs that have specifically inserted into the target sequence are recognized by the primers and amplified. The resulting TNs are cloned, the ends recombined, and the process performed recursively until the optimal level of specificity has been obtained.

To screen for reduced specificity of insertion, a library of inverted repeat sequences, e.g., in the context of a TN, or vector incorporating a TN, is delivered to a target cell population. Cells are then selected for insertion of the TN, for example by growing in the presence of a drug for which the TN carries a resistance gene. The cellular DNA is isolated and cleaved with a restriction enzyme outside the TN. The cleaved DNA is then size fractionated, e.g., by agarose gel electrophoresis. The more specific the target

site of insertion, the smaller the variation in the size distribution of the cleaved integration products. For example, a TN with a strict requirement for a specific target sequence exhibits a single band, or a few bands corresponding to the precise number of perfect matches in the cell's DNA. In contrast, a TN with low sequence specificity for integration exhibits a broad spectrum in its size distribution, e.g., a smear. TNs from cells having insertions in a distribution of pathways are amplified by the PCR, cloned, recombined, and the process is repeated until the desired level of specificity/randomness is detected.

Copy number

IS/TNs range in the number of integrated copies found in each cell. While the exact determinant of copy number is unknown, it is likely that the inverted repeats influence this property. Thus, a library ISs or TNs incorporating diversified, e.g., shuffled, inverted repeats can be screened for a change in cellular copy number. A library of TN:inverted repeats (as described above) including a gene for which copy number is quantitatively detectable, e.g., kanamycin resistance, is prepared. The library is delivered to a population of cells, and the cells are selected for resistance to increasing concentrations of kanamycin. The TNs from highly resistant cells are amplified by PCR, recombined, and the process is repeated until sufficient resistance and, thus, TN copy number is obtained. Total TN copy number and distribution within the cell can be assessed by genomic southern blot analysis using the TN as a probe.

Host adaptation

Since most genomes contain resident ISs and TNs, there are also resident transposases. Diversification, e.g., by shuffling, of the inverted repeats can lead to inverted repeat sequences recognized by these resident transposases. This provides one approach to adapting an IS or TN to a new host cell: adapting the inverted repeats to the transposases already residing in the target cell. A library of mini-TNs, i.e., transposons lacking an encoded transposase, of differing inverted repeats containing a selectable marker is delivered to a population of cells believed to possess resident transposases. The cells are selected for integration of a TN, e.g., by selection of the incorporated marker. The total number of selected cells from the library is compared to that obtained from a population of cells receiving a control, e.g., a TN having a parental set of inverted repeats. An increase in the presence of integrated TNs indicates enhanced transposition as a result of resident transposases that recognize variant inverted repeats generated by the

diversification process(es). TNs from the selected cells are amplified by PCR, recombined, and the process is repeated until the desired transposition frequency is obtained. Transposition as opposed to homologous recombination is confirmed by identification of integration sites by sequencing outward from the inserted TNs.

Increased efficiency of transposition

In addition, a library of variant, e.g., shuffled inverted repeats, e.g., TNs incorporating shuffled inverted repeats can be screened for variants that are more efficiently recombined by a particular transposase, i.e., the variants can be screened for hyper-transposable elements. To identify hyper-transposable elements, cells transformed with a TN library are selected for insertions at different periods of time after transformation. Cells that obtain TN insertions at a time point that is earlier than those transformed with the wild-type TN likely transpose with greater efficiency. These hyper-transposons are amplified from the selected cells, and the process is repeated until the transposition frequency has reached a desired level.

Transposases

Like the inverted repeats, transposases also affect the sequence specificity, the host adaptation, and the recombination efficiency of an IS or TN. Transposases can be found as single or multiple open reading frames. Many are encoded by two overlapping open reading frames such that during translation the two proteins are fused as a single polypeptide. In some cases the two open reading frames are translated both as separate proteins as well as a fusion protein. In some cases one can bind the inverted repeat sequence and inhibit the binding of the active transposase, thus, acting as a regulator, i.e., a trans-dominant regulator, of the transposase. Diversifying, e.g., by shuffling, sequences that encode transposases can be used to improve many of the same IS and TN properties as described above for the inverted repeats. Diversified transposases can be screened for recombination site specificity, i.e., more specific or more random, host adaptation, hyper-recombination, cell copy number, and the ability to mobilize other ISs and TNs within a host cell in which the transposase is expressed. Hyper-recombinogenic transposases expressed in a cell can be used to catalyze IS and TN mediated rearrangement of the cells genome, thus providing a powerful method of creating diversity within a cell population. The screens and selections described previously for site-selectivity, copy number, strain adaptability, transposition frequency, etc, can be carried out as described in the previous section.

Targeted insertion into a chromosome

ISs and TNs that undergo site specific integration do so by transposase assisted recombination. Although formally considered non-homologous recombination, the process is largely directed by a limited homology between the inverted repeats and a chomosomal insertion site. Homologous recombination between such limited regions of homology is mediated by the action of the transposase. Transposases that are evolved to work with specifically designed ("designer") inverted repeats, can be used to direct gene(s)/sequences/libraries flanked by the designer inverted repeats to specific chromosomal locations. This simple approach for targeting genes to the chromosome provides many advantages over current systems such as suicide delivery vectors. One application is to deliver fragment libraries into chromosomal expression vectors, i.e., just down stream of specific promoter or operator sequences. For example, a transposase can be evolved to target a transposon having designer inverted repeats corresponding to a specific chromosomal sequence. The resulting integration places the TN and the DNA fragments between the flanking repeats to a sequence specific locale. This process resembles gene replacement by homologous recombination rather than that typically catalyzed by a transposase. One application is the construction of a chromosomal expression cassette into which one can target any DNA, e.g., a gene of interest, to be expressed (chromosomal expression is preferred in industrial applications since it avoids the issues of plasmid loss and instability). The evolved TN/transposase system provides the tools to deliver any gene of interest to the chromosomal expression cassette such that the DNA is properly expressed. Such an approach obviates the need to carry out two steps of recombination as is required for classic gene replacement, such as that employing suicide vectors.

Integration into non-supercoiled DNA

Many transposable elements, and their transposases, e.g., the TN5 transposase, as well as their hyper-recombinogenic variants, mediate integration into supercoiled DNA with much higher efficiency than they mediate integration into non-supercoiled or relaxed, e.g., linear, DNA. As purified DNA, e.g., purified genomic DNA, is typically sheared, it is not supercoiled. Thus, the efficiency of transposition mediated by such transposases, e.g., the TN5 transposase, is not optimal. To improve the efficiency with which a transposase promotes integration into non-supercoiled, i.e., relaxed, DNA, extracts of host cells, such as *B. subtilis*, expressing variant transposases are incubated

with a mini-TN carrying a drug resistance cassette and cellular genomic DNA, under conditions suitable for transposition, e.g., in the presence of Mg^{2+}. Samples of the incubation are then transformed into host cells, e.g., *Bacillus* host cells, and the cells are screened for resistance conferred by the drug resistance marker. Alternatively, extracts from cells expressing variants can be incubated with a transposon and a single linear fragment of "recipient" DNA. Pooled samples are separated by electrophoresis and an increase in the molecular weight of the recipient Dna due to transposon integration is detected. In either case, samples expressing transposases resulting in integration into non-supercoiled DNA are isolated, e.g., by deconvolution of the samples, and can be further improved as desired.

In vitro transposition

Isolated transposases have been found to catalyze recombination between polynucleotide substrates in vitro. In particular, a variant form of TN5 has been proposed to efficiently mediate recombination between a polynucleotide having 19-bp TN5 outer end recognition sequences and a target polynucleotide (*see*, e.g., US patent No. 5,965,443 "System for in vitro Transposition" to Reznikoff et al., issued October 12, 1999, and US patent No. 5,948,622 "System for in vitro Transposition" to Reznikoff et al. issued September 7, 1999). The present invention can be used to evolve a wide variety of transposases that mediate transposition between DNA molecules in an acellular reaction mix. For example, acellular reaction mixes, each having a donor polynucleotide with transposase recognition sequences (e.g., inverted or end repeats), a target polynucleotide with which the donor can recombine, and a variant transposase expressed from a library of transposase encoding sequences or transposable elements are evaluated for frequency of recombination, e.g., by detecting a size difference between the donor, target, and recombined or "transposed" product by agarose gel electrophoresis. Library members can be evaluated singly or in pools.

Transposases with increased activity are useful, e.g., in the context of whole genome shuffling, as mediators of genetic change in cells. Improved transposases bind polynucleotides, e.g., having a gene of interest such as a marker, flanked by the appropriate recognition sequence. The complex, or "transposome" can be isolated, conveniently stored and handled, and subsequently introduced, e.g., by electroporation, into a cell of choice where the transposome effectively mediates genetic recombination. The result of the transposome mediated recombination is to introduce the donor

polynucleotide at, e.g., essentially random, locations in the genome creating a library of insertional mutant cells with a variety of structural and regulatory alterations. Such libraries are optionally screened for desired phenotypes. One such method is proposed in PCT Application No. WO 00/17343 by Reznikoff et. al., "Method for Making Insertional Mutations," published March 30, 2000.

Multi-component formats

ISs and TNs range in size from less that 1000 base pairs (ISs) to greater than 60 kb (TNs). In some cases, the properties of an individual IS or TN are not solely a property of the inverted repeat or the transposase, but rather are a holistic property of the IS or TN. Thus complete ISs and TNs can be diversified, e.g., by shuffling, and screened for any of the properties described above. For example, the size of internal DNA that can be effectively mobilized by an IS or TN is an important property with respect to its use as a vector. For the application of TN mediated whole genome shuffling, it is desirable to deliver and mobilize TNs carrying large gDNA fragments. Evolving an IS and/or TN to efficiently mobilize DNA fragments of a desired size is thus a preferred application. A fragment of DNA of desired size containing a gene for which there is a selection is cloned within a library of TNs. The library is delivered to a population of cells, and cells having insertions are selected. TNs from the selected cells are amplified by the PCR. The amplified population is separated by agarose gel electrophoresis and those having a molecular weight corresponding to a TN maintaining the complete inserted DNA are isolated, recombined, and reevaluated. This process is repeated until a TN capable of stably carrying DNA of the desired size is obtained.

DIRECTED EVOLUTION OF TRANSPOSABLE ELEMENTS

A variety of diversity generating protocols are available and described in the art. The procedures can be used separately, and/or in combination to produce one or more variants of a nucleic acid or set of nucleic acids, as well variants of encoded proteins. Individually and collectively, these procedures provide robust, widely applicable ways of generating diversified nucleic acids and sets of nucleic acids (including, e.g., nucleic acid libraries) useful, e.g., for the engineering or directed evolution of nucleic acids, proteins, pathways, cells and/or organisms with new and/or improved characteristics.

While distinctions and classifications are made in the course of the ensuing discussion for clarity, it will be appreciated that the techniques are often not mutually exclusive. Indeed, the various methods can be used singly or in combination, in parallel or in series, to access diverse sequence variants.

The result of any of the diversity generating procedures described herein can be the generation of one or more nucleic acids, which can be selected or screened for nucleic acids with or which confer desirable properties, or that encode proteins with or which confer desirable properties. Following diversification by one or more of the methods herein, or otherwise available to one of skill, any nucleic acids that are produced can be selected for a desired activity or property, e.g. transposable elements with improved in vivo or in vitro transposition efficiency, integration specificity, copy number, host specificity, etc. This can include identifying any activity that can be detected, for example, in an automated or automatable format, by any of the assays in the art, e.g., as described above. A variety of related (or even unrelated) properties can be evaluated, in serial or in parallel, at the discretion of the practitioner.

Descriptions of a variety of diversity generating procedures for producing modified transposable element nucleic acid sequences are found in the following publications and the references cited therein: Soong, N. et al. (2000) "Molecular breeding of viruses" Nat Genet 25(4):436-439; Stemmer, et al. (1999) "Molecular breeding of viruses for targeting and other clinical properties" Tumor Targeting 4:1-4; Ness et al. (1999) "DNA Shuffling of subgenomic sequences of subtilisin" Nature Biotechnology 17:893-896; Chang et al. (1999) "Evolution of a cytokine using DNA family shuffling" Nature Biotechnology 17:793-797; Minshull and Stemmer (1999) "Protein evolution by molecular breeding" Current Opinion in Chemical Biology 3:284-290; Christians et al. (1999) "Directed evolution of thymidine kinase for AZT phosphorylation using DNA family shuffling" Nature Biotechnology 17:259-264; Crameri et al. (1998) "DNA shuffling of a family of genes from diverse species accelerates directed evolution" Nature 391:288-291; Crameri et al. (1997) "Molecular evolution of an arsenate detoxification pathway by DNA shuffling," Nature Biotechnology 15:436-438; Zhang et al. (1997) "Directed evolution of an effective fucosidase from a galactosidase by DNA shuffling and screening" Proc. Natl. Acad. Sci. USA 94:4504-4509; Patten et al. (1997) "Applications of DNA Shuffling to Pharmaceuticals and Vaccines" Current Opinion in Biotechnology 8:724-733; Crameri et al. (1996) "Construction and evolution of antibody-phage libraries

by DNA shuffling" Nature Medicine 2:100-103; Crameri et al. (1996) "Improved green fluorescent protein by molecular evolution using DNA shuffling" Nature Biotechnology 14:315-319; Gates et al. (1996) "Affinity selective isolation of ligands from peptide libraries through display on a lac repressor 'headpiece dimer'" Journal of Molecular Biology 255:373-386; Stemmer (1996) "Sexual PCR and Assembly PCR" In: The Encyclopedia of Molecular Biology. VCH Publishers, New York. pp.447-457; Crameri and Stemmer (1995) "Combinatorial multiple cassette mutagenesis creates all the permutations of mutant and wildtype cassettes" BioTechniques 18:194-195; Stemmer et al., (1995) "Single-step assembly of a gene and entire plasmid form large numbers of oligodeoxy-ribonucleotides" Gene, 164:49-53; Stemmer (1995) "The Evolution of Molecular Computation" Science 270: 1510; Stemmer (1995) "Searching Sequence Space" Bio/Technology 13:549-553; Stemmer (1994) "Rapid evolution of a protein in vitro by DNA shuffling" Nature 370:389-391; and Stemmer (1994) "DNA shuffling by random fragmentation and reassembly: In vitro recombination for molecular evolution." Proc. Natl. Acad. Sci. USA 91:10747-10751.

Mutational methods of generating diversity include, for example, site-directed mutagenesis (Ling et al. (1997) "Approaches to DNA mutagenesis: an overview" Anal Biochem. 254(2): 157-178; Dale et al. (1996) "Oligonucleotide-directed random mutagenesis using the phosphorothioate method" Methods Mol. Biol. 57:369-374; Smith (1985) "In vitro mutagenesis" Ann. Rev. Genet. 19:423-462; Botstein & Shortle (1985) "Strategies and applications of in vitro mutagenesis" Science 229:1193-1201; Carter (1986) "Site-directed mutagenesis" Biochem. J. 237:1-7; and Kunkel (1987) "The efficiency of oligonucleotide directed mutagenesis" in Nucleic Acids & Molecular Biology (Eckstein, F. and Lilley, D.M.J. eds., Springer Verlag, Berlin)); mutagenesis using uracil containing templates (Kunkel (1985) "Rapid and efficient site-specific mutagenesis without phenotypic selection" Proc. Natl. Acad. Sci. USA 82:488-492; Kunkel et al. (1987) "Rapid and efficient site-specific mutagenesis without phenotypic selection" Methods in Enzymol. 154, 367-382; and Bass et al. (1988) "Mutant Trp repressors with new DNA-binding specificities" Science 242:240-245); oligonucleotide-directed mutagenesis (Methods in Enzymol. 100: 468-500 (1983); Methods in Enzymol. 154: 329-350 (1987); Zoller & Smith (1982) "Oligonucleotide-directed mutagenesis using M13-derived vectors: an efficient and general procedure for the production of point mutations in any DNA fragment" Nucleic Acids Res. 10:6487-6500; Zoller & Smith

(1983) "Oligonucleotide-directed mutagenesis of DNA fragments cloned into M13 vectors" Methods in Enzymol. 100:468-500; and Zoller & Smith (1987) "Oligonucleotide-directed mutagenesis: a simple method using two oligonucleotide primers and a single-stranded DNA template" Methods in Enzymol. 154:329-350); phosphorothioate-modified DNA mutagenesis (Taylor et al. (1985) "The use of phosphorothioate-modified DNA in restriction enzyme reactions to prepare nicked DNA" Nucl. Acids Res. 13: 8749-8764; Taylor et al. (1985) "The rapid generation of oligonucleotide-directed mutations at high frequency using phosphorothioate-modified DNA" Nucl. Acids Res. 13: 8765-8787 (1985); Nakamaye & Eckstein (1986) "Inhibition of restriction endonuclease Nci I cleavage by phosphorothioate groups and its application to oligonucleotide-directed mutagenesis" Nucl. Acids Res. 14: 9679-9698; Sayers et al. (1988) "Y-T Exonucleases in phosphorothioate-based oligonucleotide-directed mutagenesis" Nucl. Acids Res. 16:791-802; and Sayers et al. (1988) "Strand specific cleavage of phosphorothioate-containing DNA by reaction with restriction endonucleases in the presence of ethidium bromide" Nucl. Acids Res. 16: 803-814); mutagenesis using gapped duplex DNA (Kramer et al. (1984) "The gapped duplex DNA approach to oligonucleotide-directed mutation construction" Nucl. Acids Res. 12: 9441-9456; Kramer & Fritz (1987) Methods in Enzymol. "Oligonucleotide-directed construction of mutations via gapped duplex DNA" 154:350-367; Kramer et al. (1988) "Improved enzymatic in vitro reactions in the gapped duplex DNA approach to oligonucleotide-directed construction of mutations" Nucl. Acids Res. 16: 7207; and Fritz et al. (1988) "Oligonucleotide-directed construction of mutations: a gapped duplex DNA procedure without enzymatic reactions in vitro" Nucl. Acids Res. 16: 6987-6999).

Additional suitable methods include point mismatch repair (Kramer et al. (1984) "Point Mismatch Repair" Cell 38:879-887), mutagenesis using repair-deficient host strains (Carter et al. (1985) "Improved oligonucleotide site-directed mutagenesis using M13 vectors" Nucl. Acids Res. 13: 4431-4443; and Carter (1987) "Improved oligonucleotide-directed mutagenesis using M13 vectors" Methods in Enzymol. 154: 382-403), deletion mutagenesis (Eghtedarzadeh & Henikoff (1986) "Use of oligonucleotides to generate large deletions" Nucl. Acids Res. 14: 5115), restriction-selection and restriction-purification (Wells et al. (1986) "Importance of hydrogen-bond formation in stabilizing the transition state of subtilisin" Phil. Trans. R. Soc. Lond. A 317: 415-423), mutagenesis by total gene synthesis (Nambiar et al. (1984) "Total synthesis and cloning

of a gene coding for the ribonucleasc S protein" Science 223: 1299-1301; Sakamar and Khorana (1988) "Total synthesis and expression of a gene for the a-subunit of bovine rod outer segment guanine nucleotide-binding protein (transducin)" Nucl. Acids Res. 14: 6361-6372; Wells et al. (1985) "Cassette mutagenesis: an efficient method for generation of multiple mutations at defined sites" Gene 34:315-323; and Grundström et al. (1985) "Oligonucleotide-directed mutagenesis by microscale 'shot-gun' gene synthesis" Nucl. Acids Res. 13: 3305-3316), double-strand break repair (Mandecki (1986) "Oligonucleotide-directed double-strand break repair in plasmids of *Escherichia coli*: a method for site-specific mutagenesis" Proc. Natl. Acad. Sci. USA, 83:7177-7181; and Arnold (1993) "Protein engineering for unusual environments" Current Opinion in Biotechnology 4:450-455). Additional details on many of the above methods can be found in Methods in Enzymology Volume 154, which also describes useful controls for trouble-shooting problems with various mutagenesis methods.

Additional details regarding various diversity generating methods can be found in the following U.S. patents, PCT publications and applications, and EPO publications: U.S. Pat. No. 5,605,793 to Stemmer (February 25, 1997), "Methods for In Vitro Recombination;" U.S. Pat. No. 5,811,238 to Stemmer et al. (September 22, 1998) "Methods for Generating Polynucleotides having Desired Characteristics by Iterative Selection and Recombination;" U.S. Pat. No. 5,830,721 to Stemmer et al. (November 3, 1998), "DNA Mutagenesis by Random Fragmentation and Reassembly;" U.S. Pat. No. 5,834,252 to Stemmer, et al. (November 10, 1998) "End-Complementary Polymerase Reaction;" U.S. Pat. No. 5,837,458 to Minshull, et al. (November 17, 1998), "Methods and Compositions for Cellular and Metabolic Engineering;" WO 95/22625, Stemmer and Crameri, "Mutagenesis by Random Fragmentation and Reassembly;" WO 96/33207 by Stemmer and Lipschutz "End Complementary Polymerase Chain Reaction;" WO 97/20078 by Stemmer and Crameri "Methods for Generating Polynucleotides having Desired Characteristics by Iterative Selection and Recombination;" WO 97/35966 by Minshull and Stemmer, "Methods and Compositions for Cellular and Metabolic Engineering;" WO 99/41402 by Punnonen et al. "Targeting of Genetic Vaccine Vectors;" WO 99/41383 by Punnonen et al. "Antigen Library Immunization;" WO 99/41369 by Punnonen et al. "Genetic Vaccine Vector Engineering;" WO 99/41368 by Punnonen et al. "Optimization of Immunomodulatory Properties of Genetic Vaccines;" EP 752008 by Stemmer and Crameri, "DNA Mutagenesis by Random Fragmentation and Reassembly;"

EP 0932670 by Stemmer "Evolving Cellular DNA Uptake by Recursive Sequence Recombination;" WO 99/23107 by Stemmer et al., "Modification of Virus Tropism and Host Range by Viral Genome Shuffling;" WO 99/21979 by Apt et al., "Human Papillomavirus Vectors;" WO 98/31837 by del Cardayre et al. "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination;" WO 98/27230 by Patten and Stemmer, "Methods and Compositions for Polypeptide Engineering;" WO 98/27230 by Stemmer et al., "Methods for Optimization of Gene Therapy by Recursive Sequence Shuffling and Selection," WO 00/00632, "Methods for Generating Highly Diverse Libraries," WO 00/09679, "Methods for Obtaining in Vitro Recombined Polynucleotide Sequence Banks and Resulting Sequences," WO 98/42832 by Arnold et al., "Recombination of Polynucleotide Sequences Using Random or Defined Primers," WO 99/29902 by Arnold et al., "Method for Creating Polynucleotide and Polypeptide Sequences," WO 98/41653 by Vind, "An in Vitro Method for Construction of a DNA Library," WO 98/41622 by Borchert et al., "Method for Constructing a Library Using DNA Shuffling," and WO 98/42727 by Pati and Zarling, "Sequence Alterations using Homologous Recombination;" WO 00/18906 by Patten et al., "Shuffling of Codon-Altered Genes;" WO 00/04190 by del Cardayre et al. "Evolution of Whole Cells and Organisms by Recursive Recombination;" WO 00/42561 by Crameri et al., "Oligonucleotide Mediated Nucleic Acid Recombination;" WO 00/42559 by Selifonov and Stemmer "Methods of Populating Data Structures for Use in Evolutionary Simulations;" WO 00/42560 by Selifonov et al., "Methods for Making Character Strings, Polynucleotides & Polypeptides Having Desired Characteristics;" WO 01/23401 by Welch et al., "Use of Codon-Varied Oligonucleotide Synthesis for Synthetic Shuffling;" and PCT/US01/06775 "Single-Stranded Nucleic Acid Template-Mediated Recombination and Nucleic Acid Fragment Isolation" by Affholter.

In brief, several different general classes of sequence modification methods, such as mutation, recombination, etc. are applicable to the generation of transposable elements (e.g., transposons, insertion sequences, and their components) with desired properties, and set forth, e.g., in the references above.

The following exemplify some of the different types of preferred formats for diversity generation in the context of the present invention, including, e.g., certain recombination based diversity generation formats.

Nucleic acids can be recombined in vitro by any of a variety of techniques discussed in the references above, including e.g., DNAse digestion of nucleic acids to be recombined followed by ligation and/or PCR reassembly of the nucleic acids. For example, sexual PCR mutagenesis can be used in which random (or pseudo random, or even non-random) fragmentation of the DNA molecule is followed by recombination, based on sequence similarity, between DNA molecules with different but related DNA sequences, in vitro, followed by fixation of the crossover by extension in a polymerase chain reaction. This process and many process variants is described in several of the references above, e.g., in Stemmer (1994) Proc. Natl. Acad. Sci. USA 91:10747-10751. Thus, transposable elements with desired properties, such as increased transposase activity, increased in vitro transposition activity, altered host specificity, targeted insertion, and the like, can be produced by in vitro recombination procedures.

Similarly, nucleic acids can be recursively recombined in vivo, e.g., by allowing recombination to occur between nucleic acids in cells. Many such in vivo recombination formats are set forth in the references noted above. Such formats optionally provide direct recombination between nucleic acids of interest, or provide recombination between vectors, viruses, plasmids, etc., comprising the nucleic acids of interest, as well as other formats. Details regarding such procedures are found in the references noted above. Thus, in vivo recombination procedures can be employed to recombine and select transposable elements with improved properties.

Whole genome recombination methods can also be used in which whole genomes of cells or other organisms are recombined, optionally including spiking of the genomic recombination mixtures with desired library components (e.g., genes corresponding to the pathways of the present invention). These methods have many applications, including those in which the identity of a target gene is not known. Details on such methods are found, e.g., in WO 98/31837 by del Cardayre et al. "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination;" and in, e.g., WO 00/04190 by del Cardayre et al., also entitled "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination." Such methods can be used to generate variant transposable elements with new and improved characteristics, e.g., by recombining genomes harboring one or more transposable element, and, optionally by introducing into such cells, additional sequences derived from libraries of nucleic acids, e.g., comprising components of one or more transposable element.

Synthetic recombination methods can also be used, in which oligonucleotides corresponding to targets of interest are synthesized and reassembled in PCR or ligation reactions which include oligonucleotides which correspond to more than one parental nucleic acid, thereby generating new recombined nucleic acids. Oligonucleotides can be made by standard nucleotide addition methods, or can be made, e.g., by tri-nucleotide synthetic approaches. Details regarding such approaches are found in the references noted above, including, e.g., WO 00/42561 by Crameri et al., "Olgonucleotide Mediated Nucleic Acid Recombination;" WO 01/23401 by Welch et al., "Use of Codon-Varied Oligonucleotide Synthesis for Synthetic Shuffling;" WO 00/42560 by Selifonov et al., "Methods for Making Character Strings, Polynucleotides and Polypeptides Having Desired Characteristics;" and WO 00/42559 by Selifonov and Stemmer "Methods of Populating Data Structures for Use in Evolutionary Simulations."

In silico methods of recombination can be effected in which genetic algorithms are used in a computer to recombine sequence strings which correspond to homologous (or even non-homologous) nucleic acids. The resulting recombined sequence strings are optionally converted into nucleic acids by synthesis of nucleic acids which correspond to the recombined sequences, e.g., in concert with oligonucleotide synthesis/ gene reassembly techniques. This approach can generate random, partially random or designed variants. Many details regarding in silico recombination, including the use of genetic algorithms, genetic operators and the like in computer systems, combined with generation of corresponding nucleic acids (and/or proteins), as well as combinations of designed nucleic acids and/or proteins (e.g., based on cross-over site selection) as well as designed, pseudo-random or random recombination methods are described in WO 00/42560 by Selifonov et al., "Methods for Making Character Strings, Polynucleotides and Polypeptides Having Desired Characteristics" and WO 00/42559 by Selifonov and Stemmer "Methods of Populating Data Structures for Use in Evolutionary Simulations." Extensive details regarding in silico recombination methods are found in these applications. This methodology is generally applicable to the present invention in providing for recombination of transposable elements and their components in silico and/ or the generation of corresponding nucleic acids or proteins.

Many methods of accessing natural diversity, e.g., by hybridization of diverse nucleic acids or nucleic acid fragments to single-stranded templates, followed by polymerization and/or ligation to regenerate full-length sequences, optionally followed by

degradation of the templates and recovery of the resulting modified nucleic acids can be similarly used. In one method employing a single-stranded template, the fragment population derived from the genomic library(ies) is annealed with partial, or, often approximately full length ssDNA or RNA corresponding to the opposite strand. Assembly of complex chimeric genes from this population is then mediated by nuclease-base removal of non-hybridizing fragment ends, polymerization to fill gaps between such fragments and subsequent single stranded ligation. The parental polynucleotide strand can be removed by digestion (e.g., if RNA or uracil-containing), magnetic separation under denaturing conditions (if labeled in a manner conducive to such separation) and other available separation/purification methods. Alternatively, the parental strand is optionally co-purified with the chimeric strands and removed during subsequent screening and processing steps. Additional details regarding this approach are found, e.g., in "Single-Stranded Nucleic Acid Template-Mediated Recombination and Nucleic Acid Fragment Isolation" by Affholter, PCT/US01/06775.

In another approach, single-stranded molecules are converted to double-stranded DNA (dsDNA) and the dsDNA molecules are bound to a solid support by ligand-mediated binding. After separation of unbound DNA, the selected DNA molecules are released from the support and introduced into a suitable host cell to generate a library enriched sequences which hybridize to the probe. A library produced in this manner provides a desirable substrate for further diversification using any of the procedures described herein.

Any of the preceding general recombination formats can be practiced in a reiterative fashion (e.g., one or more cycles of mutation/recombination or other diversity generation methods, optionally followed by one or more selection methods) to generate a more diverse set of recombinant nucleic acids.

Mutagenesis employing polynucleotide chain termination methods have also been proposed (*see* e.g., U.S. Patent No. 5,965,408, "Method of DNA reassembly by interrupting synthesis" to Short, and the references above), and can be applied to the present invention. In this approach, double stranded DNAs corresponding to one or more genes sharing regions of sequence similarity are combined and denatured, in the presence or absence of primers specific for the gene. The single stranded polynucleotides are then annealed and incubated in the presence of a polymerase and a chain terminating reagent (e.g., ultraviolet, gamma or X-ray irradiation; ethidium bromide or other intercalators;

DNA binding proteins, such as single strand binding proteins, transcription activating factors, or histones; polycyclic aromatic hydrocarbons; trivalent chromium or a trivalent chromium salt; or abbreviated polymerization mediated by rapid thermocycling; and the like), resulting in the production of partial duplex molecules. The partial duplex molecules, e.g., containing partially extended chains, are then denatured and reannealed in subsequent rounds of replication or partial replication resulting in polynucleotides which share varying degrees of sequence similarity and which are diversified with respect to the starting population of DNA molecules. Optionally, the products, or partial pools of the products, can be amplified at one or more stages in the process. Polynucleotides produced by a chain termination method, such as described above, are suitable substrates for any other described recombination format.

Diversity also can be generated in nucleic acids or populations of nucleic acids using a recombinational procedure termed "incremental truncation for the creation of hybrid enzymes" ("ITCHY") described in Ostermeier et al. (1999) "A combinatorial approach to hybrid enzymes independent of DNA homology" Nature Biotech 17:1205. This approach can be used to generate an initial a library of variants which can optionally serve as a substrate for one or more in vitro or in vivo recombination methods. See, also, Ostermeier et al. (1999) "Combinatorial Protein Engineering by Incremental Truncation," Proc. Natl. Acad. Sci. USA, 96: 3562-67; Ostermeier et al. (1999), "Incremental Truncation as a Strategy in the Engineering of Novel Biocatalysts," Biological and Medicinal Chemistry, 7: 2139-44.

Mutational methods which result in the alteration of individual nucleotides or groups of contiguous or non-contiguous nucleotides can be favorably employed to introduce nucleotide diversity into transposable elements and their components. Many mutagenesis methods are found in the above-cited references; additional details regarding mutagenesis methods can be found in following, which can also be applied to the present invention.

For example, error-prone PCR can be used to generate nucleic acid variants. Using this technique, PCR is performed under conditions where the copying fidelity of the DNA polymerase is low, such that a high rate of point mutations is obtained along the entire length of the PCR product. Examples of such techniques are found in the references above and, e.g., in Leung et al. (1989) Technique 1:11-15 and Caldwell et al. (1992) PCR Methods Applic. 2:28-33. Similarly, assembly PCR can be

used, in a process which involves the assembly of a PCR product from a mixture of small DNA fragments. A large number of different PCR reactions can occur in parallel in the same reaction mixture, with the products of one reaction priming the products of another reaction.

Oligonucleotide directed mutagenesis can be used to introduce site-specific mutations in a nucleic acid sequence of interest. Examples of such techniques are found in the references above and, e.g., in Reidhaar-Olson et al. (1988) Science, 241:53-57. Similarly, cassette mutagenesis can be used in a process that replaces a small region of a double stranded DNA molecule with a synthetic oligonucleotide cassette that differs from the native sequence. The oligonucleotide can contain, e.g., completely and/or partially randomized native sequence(s).

Recursive ensemble mutagenesis is a process in which an algorithm for protein mutagenesis is used to produce diverse populations of phenotypically related mutants, members of which differ in amino acid sequence. This method uses a feedback mechanism to monitor successive rounds of combinatorial cassette mutagenesis. Examples of this approach are found in Arkin & Youvan (1992) Proc. Natl. Acad. Sci. USA 89:7811-7815.

Exponential ensemble mutagenesis can be used for generating combinatorial libraries with a high percentage of unique and functional mutants. Small groups of residues in a sequence of interest are randomized in parallel to identify, at each altered position, amino acids which lead to functional proteins. Examples of such procedures are found in Delegrave & Youvan (1993) Biotechnology Research 11:1548-1552.

In vivo mutagenesis can be used to generate random mutations in any cloned DNA of interest by propagating the DNA, e.g., in a strain of *E. coli* that carries mutations in one or more of the DNA repair pathways. These "mutator" strains have a higher random mutation rate than that of a wild-type parent. Propagating the DNA in one of these strains will eventually generate random mutations within the DNA. Such procedures are described in the references noted above.

Other procedures for introducing diversity into a genome, e.g. a bacterial, fungal, animal or plant genome can be used in conjunction with the above described and/or referenced methods. For example, in addition to the methods above, techniques have been proposed which produce nucleic acid multimers suitable for transformation

into a variety of species (*see*, e.g., Schellenberger U.S. Patent No. 5,756,316 and the references above). Transformation of a suitable host with such multimers, consisting of genes that are divergent with respect to one another, (e.g., derived from natural diversity or through application of site directed mutagenesis, error prone PCR, passage through mutagenic bacterial strains, and the like), provides a source of nucleic acid diversity for DNA diversification, e.g., by an in vivo recombination process as indicated above.

Alternatively, a multiplicity of monomeric polynucleotides sharing regions of partial sequence similarity can be transformed into a host species and recombined in vivo by the host cell. Subsequent rounds of cell division can be used to generate libraries, members of which, include a single, homogenous population, or pool of monomeric polynucleotides. Alternatively, the monomeric nucleic acid can be recovered by standard techniques, e.g., PCR and/or cloning, and recombined in any of the recombination formats, including recursive recombination formats, described above.

Methods for generating multispecies expression libraries have been described (in addition to the reference noted above, *see*, e.g., Peterson et al. (1998) U.S. Pat. No. 5,783,431 "Methods for Generating and Screening Novel Metabolic Pathways," and Thompson, et al. (1998) U.S. Pat. No. 5,824,485 Methods for Generating and Screening Novel Metabolic Pathways) and their use to identify protein activities of interest has been proposed (In addition to the references noted above, *see*, Short (1999) U.S. Pat. No. 5,958,672 "Protein Activity Screening of Clones Having DNA from Uncultivated Microorganisms"). Multispecies expression libraries include, in general, libraries comprising cDNA or genomic sequences from a plurality of species or strains, operably linked to appropriate regulatory sequences, in an expression cassette. The cDNA and/or genomic sequences are optionally randomly ligated to further enhance diversity. The vector can be a shuttle vector suitable for transformation and expression in more than one species of host organism, e.g., bacterial species, eukaryotic cells. In some cases, the library is biased by preselecting sequences which encode a protein of interest, or which hybridize to a nucleic acid of interest. Any such libraries can be provided as substrates for any of the methods herein described.

The above described procedures have been largely directed to increasing nucleic acid and/ or encoded protein diversity. However, in many cases, not all of the diversity is useful, e.g., functional, and contributes merely to increasing the background of variants that must be screened or selected to identify the few favorable variants. In

some applications, it is desirable to preselect or prescreen libraries (e.g., an amplified library, a genomic library, a cDNA library, a normalized library, etc.) or other substrate nucleic acids prior to diversification, e.g., by recombination-based mutagenesis procedures, or to otherwise bias the substrates towards nucleic acids that encode functional products. For example, in the case of antibody engineering, it is possible to bias the diversity generating process toward antibodies with functional antigen binding sites by taking advantage of in vivo recombination events prior to manipulation by any of the described methods. For example, recombined CDRs derived from B cell cDNA libraries can be amplified and assembled into framework regions (e.g., Jirholt et al. (1998) "Exploiting sequence space: shuffling in vivo formed complementarity determining regions into a master framework" Gene 215: 471) prior to diversifying according to any of the methods described herein.

Libraries can be biased towards nucleic acids which encode proteins with desirable enzyme activities. For example, after identifying a clone from a library which exhibits a specified activity, the clone can be mutagenized using any known method for introducing DNA alterations. A library comprising the mutagenized homologues is then screened for a desired activity, which can be the same as or different from the initially specified activity. An example of such a procedure is proposed in Short (1999) U.S. Patent No. 5,939,250 for "Production of Enzymes Having Desired Activities by Mutagenesis." Desired activities can be identified by any method known in the art. For example, WO 99/10539 proposes that gene libraries can be screened by combining extracts from the gene library with components obtained from metabolically rich cells and identifying combinations which exhibit the desired activity. It has also been proposed (e.g., WO 98/58085) that clones with desired activities can be identified by inserting bioactive substrates into samples of the library, and detecting bioactive fluorescence corresponding to the product of a desired activity using a fluorescent analyzer, e.g., a flow cytometry device, a CCD, a fluorometer, or a spectrophotometer.

Libraries can also be biased towards nucleic acids which have specified characteristics, e.g., hybridization to a selected nucleic acid probe. For example, application WO 99/10539 proposes that polynucleotides encoding a desired activity (e.g., an enzymatic activity, for example: a lipase, an esterase, a protease, a glycosidase, a glycosyl transferase, a phosphatase, a kinase, an oxygenase, a peroxidase, a hydrolase, a hydratase, a nitrilase, a transaminase, an amidase or an acylase) can be identified from

among genomic DNA sequences in the following manner. Single stranded DNA molecules from a population of genomic DNA are hybridized to a ligand-conjugated probe. The genomic DNA can be derived from either a cultivated or uncultivated microorganism, or from an environmental sample. Alternatively, the genomic DNA can be derived from a multicellular organism, or a tissue derived therefrom. Second strand synthesis can be conducted directly from the hybridization probe used in the capture, with or without prior release from the capture medium or by a wide variety of other strategies known in the art. Alternatively, the isolated single-stranded genomic DNA population can be fragmented without further cloning and used directly in, e.g., a recombination-based approach, that employs a single-stranded template, as described above.

"Non-Stochastic" methods of generating nucleic acids and polypeptides are alleged in Short "Non-Stochastic Generation of Genetic Vaccines and Enzymes" WO 00/46344. These methods, including proposed non-stochastic polynucleotide reassembly and site-saturation mutagenesis methods be applied to the present invention as well. Random or semi-random mutagenesis using doped or degenerate oligonucleotides is also described in, e.g., Arkin and Youvan (1992) "Optimizing nucleotide mixtures to encode specific subsets of amino acids for semi-random mutagenesis" Biotechnology 10:297-300; Reidhaar-Olson et al. (1991) "Random mutagenesis of protein sequences using oligonucleotide cassettes" Methods Enzymol. 208:564-86; Lim and Sauer (1991) "The role of internal packing interactions in determining the structure and stability of a protein" J. Mol. Biol. 219:359-76; Breyer and Sauer (1989) "Mutational analysis of the fine specificity of binding of monoclonal antibody 51F to lambda repressor" J. Biol. Chem. 264:13355-60); and "Walk-Through Mutagenesis" (Crea, R; US Patents 5,830,650 and 5,798,208, and EP Patent 0527809 B1.

It will readily be appreciated that any of the above described techniques suitable for enriching a library prior to diversification can also be used to screen the products, or libraries of products, produced by the diversity generating methods.

Kits for mutagenesis, library construction and other diversity generation methods are also commercially available. For example, kits are available from, e.g., Stratagene (e.g., QuickChange™ site-directed mutagenesis kit; and Chameleon™ double-stranded, site-directed mutagenesis kit), Bio/Can Scientific, Bio-Rad (e.g., using the Kunkel method described above), Boehringer Mannheim Corp., Clonetech Laboratories, DNA Technologies, Epicentre Technologies (e.g., 5 prime 3 prime kit); Genpak Inc,

Lemargo Inc, Life Technologies (Gibco BRL), New England Biolabs, Pharmacia Biotech, Promega Corp., Quantum Biotechnologies, Amersham International plc (e.g., using the Eckstein method above), and Anglian Biotechnology Ltd (e.g., using the Carter/Winter method above).

The above references provide many mutational formats, including recombination, recursive recombination, recursive mutation and combinations or recombination with other forms of mutagenesis, as well as many modifications of these formats. Regardless of the diversity generation format that is used, the nucleic acids of the invention can be recombined (with each other, or with related (or even unrelated) sequences) to produce a diverse set of recombinant nucleic acids, including, e.g., sets of homologous nucleic acids, as well as corresponding polypeptides.

Any of these or other available diversity generating methods can be combined, in any combination selected by the user, to produce nucleic acid diversity, which can be screened or selected for using any available screening or selection method to identify evolved transposable elements or TE components as described herein.

In one aspect, the present invention provides for the recursive use of any of the diversity generation methods noted above, in any combination, to evolve nucleic acids or libraries of recombinant nucleic acids that encode enzymes involved in transposition or that are transposable elements, including both cis- and trans-acting mobilization functions. In particular, as noted, the relevant nucleic acids, e.g., TNs, Iss, transposase, inverted repeats, etc., can be modified before selection, or can be selected and then recombined, or both. This process can be reiteratively repeated until a desired property in obtained.

Regardless of the diversity generating method or methods employed, identification of novel transposable elements and TE components involves one or more screening and/or selection protocol distinguishing nucleic acids encoding products with desired properties. In some instances, the desired property or characteristic relates to the nucleic acid, e.g., hybridization, amplification, or the like. However, in many cases the desired characteristic relates to a functional property conferred by the recombinant nucleic acid, e.g., inverted repeat, ORF encoding a transposase, etc, expressed in situ.

TRANSPOSABLE ELEMENTS AS VECTORS

The breeding of a population of microbes can be facilitated by the use of "mobilizable" genomic libraries that are delivered via transposable elements. In general, genomic DNA from a population of organisms is fragmented and cloned within a transposable element. This "transposable library" is then delivered to a desired host or a population of hosts, such as the original population of organisms. Delivery can be via transformation of the library on a suicide or conditionally replicative vector, e.g., by electroporation or other well-known transformation technique, or via conjugative delivery, if the library is cloned within a conjugative transposon.

There are many variations on the nature of the transposable element into which the gDNA is cloned that can alter the effectiveness of the approach. For example, the transposable element can be an insertion element, a transposon, or a conjugative transposon. These elements can be "mini-transposable elements," such that the transposition genes are removed and provided in trans. Mini-transposable elements are preferable in some cases since incorporation into the host genome is stable in the absence of transposition factors, e.g., a transposase. Once a transposon shuffled library of microorganisms has been generated, it can be screened for desired phenotypes. The sub-population resulting from the screening can then be further bred and screened using the same methodology until a desired phenotype is achieved.

One classic method of microbial strain improvement is expression cloning. This process involves cloning genomic DNA into an expression vector, and then transforming the expression library into a desired host organism. The transformants having improved properties are then identified by an appropriate screen or selection. A similar approach is accomplished using transposons. A genomic DNA library is cloned, e.g., into a transposon or mini-transposon and delivered to the chromosome of a target organism. In addition to delivering the library sequences, the transposable element delivery vehicle explores multiple insertion sites within the genome providing an additional empirical parameter than can be optimized in seeking the desired cell phenotype.

Transformants that have improved properties are then isolated. Since the sequence of the TN is known, PCR primers directed to the TN are sufficient to amplify the transposed gDNA. In one approach, each amplified gDNA is shuffled independently, and subcloned into the original TN delivery vector. The result is several libraries each

originating from the gDNA amplified from a single improved clone. These are pooled and used to transform the original host strain, with further improvements being obtained by screening.

GENERAL DELIVERY VECTORS

One goal for TN and IS mediated genome diversification, e.g., shuffling, is the delivery of libraries of DNA fragments to a population of cells such that that members of the library are stably incorporated into the genomes of the cells. A general set of delivery vectors are described that can be used for this purpose, *see*, Figures 1A-C. The vectors share several common components (Figure 1A): an origin of replication active in a convenient cloning host, a conditional origin of replication for the target cell into which the library is being delivered, markers for positive selection in both hosts, a mini-transposon (two inverted repeats surrounding a multiple-cloning site), and, optionally, a transposase that catalyzes the mobilization of the sequence contained between the inverted repeats linked to a promoter that drives the expression of the transposase in the target cell. In some alternatives, the transposase is supplied in trans on a second vector or integrated into the genome of the target cell. The vectors are preferably designed in modular fashion to facilitate adaptation to new host cells or for different applications (examples are provided in Figures 1B and 1C). It will be appreciated that the specific choices of components are not essential to the invention and that numerous sequences are available to fulfill each function recited above. The specific choices will be apparent to those of skill in the art based on the specific application under consideration. The following examples are provided as illustration not as limitation.

Origin of replication for cloning host

Origins of replication can be derived from any plasmid that replicates in a desirable host useful for molecular cloning for the project of interest. These most often will be for *E.coli*, but can also be chosen for use in other common organisms such as *bacillus*, *synechosystis*, streptomyces, cornybacterium, lactic acid bacteria, yeast, and fungi. Some examples are: ColE1 series, pACYC series (p15A), RK4, pCM595, pSa, RK6, pUB110, pE194, pG+, SLP1, pMEA100, pSAM2, pSG1, pIJ408, pIJ110, pSE101, pSE211, pAMβ1, pIP501, pAC1, pRI405, pIP612, pIP613, pIP646, pIP920, pMV103, pMV141, pSF9400, p43, pSM19035, pERL1, pSM10419, pT181, pC221, pC223, pS194, pUB112, pCW7, pHD2, pC194, pUB110, pOX6, pLS11, pTA1060, pBAA1, pBS2,

pUG1, pFTB14, pBC16, pBC1, pCB101, pLP1, pIJ101, pC30i1, pTD1, pKYM, ϕX174, pLAB1000, pWGB32, pVA380-1, pRF1, pE194, pMV158, pWV01, pSH71, pFX2, pLB4, pA1, pADB201, pKMK1, pHPK255, pSN2, pE12, pE5, pT48, pTCS1, pNE131, pIM13, pTKX14, 2 micron circle based plasmids, artificial chromosomes, etc.

Conditional origins of replication

pSA3, pE194tm, pG+tm, are all temperature sensitive replicons for Gram-positive bacteria. There are also mutants of plasmid replication origins for Gram-negative bacteria that deem those plasmids conditionally replicative. Alternatively, conditional origins suitable for maintaining episomal replication in eukaryotic hosts can be employed.

Selection markers

Markers conferring resistance to antibiotics, prototrophy to auxotrophic organisms, or resistance to toxic compounds. Some examples are: kanamycin resistance (aph3A, and others), ampicillin resistance, macrolide-lincosamine-streptogramin (MLS) resistance, as well as resistance to apramycin, spiramycin, hygromycin, chloramphenicol, tetracycline, and many other compounds.

Mini-transposon

In the context of a vector, a mini-transposon (or mini-IS) is simply the inverted repeats of a transposon or IS element flanking a sequence of DNA, most frequently a multiple-cloning site, into which a library of DNA fragments can be cloned. The inverted repeats of the transposable element used should be such that the expressed transposase on the same plasmid (or supplied in trans) recognizes them as recombination substrates. The inverted repeats and mobilization genes can originate from any TN or IS element that can function in the target host into which the mini-TN is to integrate. A partial list of possible TNs and IS elements functioning in a variety of target organisms is provided above.

Transposase

Mobilization enzymes, i.e., transposases, are, in general, one or more enzymes, including integrases, recombinase, e.g., xis, int encoded polypeptides, that catalyze the excision and integration of the mini-TN into the target host cell genome. These genes encode enzymes that recognize the inverted repeats of the mini-transposon of the vector. These can be wild-type mobilization enzymes or ones which have been optimized by directed evolution, e.g., DNA shuffling. In many circumstances, it is most convenient to supply the transposase on the same vector as the mini-transposon, thus, in

fact, supplying a transposon. In such cases, it is often preferable to locate the transposase in close proximity to the ends of the inverted repeats. The precise meaning of "close proximity" will vary from vector to vector, and can be interpreted to mean close enough to insure efficient mobilization of the mini-TN by the transposase. The requirements of the particular vector will be readily determined experimentally. In some cases this will be adjacent to one of the inverted repeats, while in other cases more relaxed requirements will be observed.

Promoter

A promoter can be any sequence of DNA that directs the constitutive or controlled expression of the down stream mobilization gene(s), e.g., transposase, int gene, xis gene, etc. These sequences, like the conditional origin of replication are often host specific, and thus are selected to function in the host into which the mini-transposon of the vector is targeted for integration. Under some circumstances, it is preferable to use an inducible promoter that can be tightly regulated by the practitioner. In other cases, constitutive or transient promoters are selected. In some cases, the promoter is selected from among the endogenous promoters of the host cell.

ACTIVATING DORMANT / LATENT TRANSPOSITION

Evolved mobilization enzymes (e.g., transposases, integrases, recombinases, etc.) of the present invention can be used to activate dormant transposition activities in prokaryotic or eukaryotic cells. For example, a cell population (comprising known or unknown transposable elements) can be transformed with a library of plasmids expressing, e.g., evolved mobilization enzymes of the present invention, preferably under the control of an inducible promoter, and the cell population screened for increased transposition frequency. The increased transposition frequency can be assessed relative to background (e.g., uninduced) transposition frequency by comparing the transposition frequency of a cell population transformed with plasmid expressing transposase to that of a cell population transformed with plasmid lacking transposase (or, if transposase is under the control of an inducible promoter, cells grown in the absence of inducer). For example, transposition frequency can be assessed by the generation of auxotrophic mutations in a cell population by comparing the number of cell colonies present in serial dilutions plated onto minimal media plates vs. rich media plates. Transposition frequency can also be assessed in cells by monitoring the appearance of knockout mutations in a

marker gene (e.g., by loss of fluorescence when the marker gene is GFP) and/or by the appearance of papillated colonies or other morphological changes. The transposable elements (e.g., IS elements) activated by the transposase can be identified by PCR-amplifying and sequencing the knocked-out selectable marker genes.

Cells comprising dormant transposable elements identified as described above are useful in developing mutator-like strains in which transposition is activated in a controlled manner, e.g., by addition (or induction) of the cognate transposase. Such inducible mutator strains are useful for *in vivo* mutagenesis applications, such as evolving cells for improved phenotypes as described herein.

TRANSPOSITION VIA INTERMEDIATE HOST

One difficulty presented by many transposable elements is the preference of the transposase for supercoiled DNA. In the absence of a transposable element vector/transposase specific for relaxed (non-supercoiled) DNA, genome diversification can be accomplished using an intermediate host organism. In the following illustrative example, transposon mediated recombination of *Bacillus* genomic DNA is accomplished using *E. coli* as an intermediate host. For example, to recombine genomic DNA between *B. subtilis* and another organism, genomic DNA (gDNA) from the two organisms is prepared (by standard methods). A *Bacillus* gDNA library is then prepared in an appropriate *E. coli* vector, such as a bacterial artificial chromosome (BAC) or other low copy number plasmid, e.g., pACYC, that can harbor DNA fragments of at least 2 kb (preferably greater than about 10 kb). A gDNA library of the other organism(s) is prepared in a mini-TN, such as the mini-TN5 of pMOD (Epicentre). The TN gDNA library is then integrated into the plasmid (BAC) gDNA library of *B. subtilis*, which is supercoiled as purified from *E. coli*. The TN library inserts throughout the plasmid gDNA library, resulting in a plasmid encoded TN-mediated recombinant genomic library. The products of this reaction are then transformed into *E. coli* to "clean up the reaction," i.e., to fill in and ligate the broken ends resulting from the insertion reaction, and screened (or selected) for the presence of the plasmid library. Plasmid DNA is then isolated from the pool of transformants harboring the selected colonies. This isolated plasmid library is the transformed into naturally competent *Bacillus*, and the *Bacillus* gDNA is incorporated into the *Bacillus* genome by homologous recombination, carrying with it any genomic DNA from the donor species that has been integrated via the transposable element vector.

The transformed cells are then screened or selected for cells having desired properties, such as acid tolerance, heat tolerance, or improved production of a desired metabolite, etc.

IMPROVED VECTORS FOR INTEGRATION INTO MAMMALIAN CELLS

Although active transposable elements are recognized in many invertebrates, and inactive remnants of transposable elements are observed in vertebrate, including mammalian cells, no naturally transposing elements are known in mammalian cells. This limits the application of this valuable tool to mammalian cells. The present invention is used to develop transposable element vectors that efficiently integrate into mammalian, including human cells. While many sequences are suitable as substrates in the generation of such a vector, one particularly attractive candidate group of sequences are the *Mariner* transposable elements. Many suchTEs are known that transpose in a broad host range, including higher eukaryotic cells. To facilitate screening of a diversified library of transposable elements for their ability to mediate integration into the genome of mammalian cells a vector incorporating from 5' to 3': a promoter; a splice donor site; a first inverted repeat; a transposase having a splice acceptor site at its upstream terminus; a selectable marker; and a second inverted repeat. An exemplary vector is illustrated in Figure 2A. The target cell population is transfected with the vector which transiently expresses the transposase from a message spliced between the splice donor and acceptor sites. When a transposase capable of mediating integration in the selected cell type is expressed, transposition of the sequences flanked by the inverted repeats into the cellular genome can occur. Cells that have integrated these sequences survive selection based on the selectable marker, e.g., neomycin resistance. Following integration the transposase is inactive due to a separation between the promoter and the coding sequence. The coding sequences can nonetheless be recovered by PCR and further recombined and selected, following reconstruction of the vector, if desired. The entire process can be performed recursively until a desired level of transposition is achieved.

TRANSPOSONS AS AGENTS OF GENOME DIVERSIFICATION

Directed evolution of whole genomes, e.g., genome shuffling, is a combination of two processes: genome diversification (e.g., intra-genome shuffling) and genome recombination (e.g., inter-genome shuffling). Transposable elements affect both

of these processes, and are employed in the present invention to accclerate whole cell evolution. Insertion sequences and transposons catalyze the structural and functional diversification of genomes by a variety of genetic phenomena. These include gene activation, inactivation, and attenuation, sequence inversion, duplication, deletion, and mobilization, homologous recombination, and other rearrangements. In nature, these events occur spontaneously and can also be induced by cellular stress, such as starvation or exposure to extreme environments. In addition, such events can be induced artificially by activating the enzymatic machinery of transposition, e.g., through activation of an inducible promoter.

IS elements, mini-IS elements, transposons, and mini-transposons are introduced into host cells using appropriate delivery vectors and transformation techniques. For example, plasmid vectors incorporating transposable elements can be introduced into the selected host cell population by any of a number of known techniques, e.g., microinjection, electroporation, agrobacterium mediated transformation, calcium phosphate precipitation, etc. Alternatively, isolated transposomes can be introduced, e.g., by electroporation, into the cells. Which technique is selected is largely a matter to be determined by the particular application and host cell type, and will be apparent to one of skill in the art.

Integration and mobilization of these elements within the genomes of the transfected cells result in the diversification of the cell population by the mechanisms described above. This diversification can be iteratively induced by either transiently expressing the transposase or by exposing the population to periodic stress. For example, an IS element known to be induced by nitrogen starvation is delivered to a population of cells on a plasmid. The cell population is then grown under nitrogen limiting conditions to induce the intra-genomic transposition of the IS element throughout the genomes of the transfected cells. The result is a diverse population of cells having different chromosomal insertions and rearrangements. An alternative is to deliver a mini-IS element, in which the transposase has been removed from within the mobile element and placed elsewhere in the genome under an inducible promoter. Upon induction, the transposase is expressed and catalyzes the mobilization of the mini-IS elements and the corresponding genomic rearrangements. The difference between these two strategies is that the mini-IS elements cannot mobilize without the transposase being induced or provided in trans. Thus, the final strains will be more stable than those having naturally inducible transposases within

the IS elements. Processes using natural IS elements or transposons access the natural mechanisms of genome plasticity, while those using the mini-IS elements and transposons are designed to accelerate and control these natural processes. Both are of value for the purpose of directed cellular evolution.

The population resulting from the IS element mediated diversification is enriched for improved variants by either screening or selection. One preferred method for the enrichment of organisms having improved environmental tolerance is to grow the population under increasingly stringent conditions in a chemostat or turbidostat. The growing populations are slowly exposed to conditions of increasing stringency, such as increased temperature or pH. Variants having improved tolerance overtake the population. It is important that conditions are not made so stringent that no cells survive or that only a single clone survives. Rather, genetic diversity within the tolerant population is maintained and selective conditions are generally such that a group of improved variants survive. This tolerant population can then be further diversified as a result of the stressful conditions naturally inducing the mobilization of the IS elements i.e., continuously adapting to the conditions imposed. Alternatively, the population can be diversified by transiently inducing the expression of a transposase after each step of increased environmental stringency. An additional strategy of enrichment is the oscillation between stringent and permissive conditions. The diverse population is gradually exposed to an environmental challenge such that a significant portion of the population is removed. The survivors are gradually returned to permissive temperature, where they further diversify (naturally or by induction), and then gradually back to conditions slightly more stringent than the previous challenge. This process is repeated recursively until the population can tolerate no further increase in challenge. At this point, the evolutionary process benefits from the recombination of genetic information between cells existing within the population, e.g., by cellular fusion, or other described methods.

The genetic information within a population of improved cells can be recombined by any of the previously described methods for whole genome recombination, e.g., shuffling. Whole genome recombination of the improved population will generate a combinatorial genetic library of cells and/or genomes having all possible combinations of the genetic rearrangements present in the improved population. Further details regarding whole genome shuffling are provided, e.g., in USSN 116,188 and PCT

publication WO 00/04190 (1/27/2000) "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination," by del Cardyre et al. filed July 15, 1999. This library is then subjected to further phenotypic enrichments and intra-genomic shuffling. The iterative process of intra-genomic shuffling enrichment, and inter-genomic shuffling is cycled until the phenotype of interest is achieved.

TRANSPOSOME MEDIATED GENOME DIVERSIFICATION

Diversification of whole genomes can also be accomplished in vitro using transposomes to mediate the recombination events. This method provides a means of efficiently recombining the genomic DNA from multiple different organisms in vitro. Large fragments of genomic DNA are recombined, e.g., shuffled, in vitro by transposase-mediated non-homologous recombination. The resulting diverse library is then delivered to a target host organism, e.g., where homologous recombination of the library with the host genome results in chromosomal variations that mimic in vivo transposition of heterologous DNA.

Genomic DNA is purified using standard procedures from various sources according to the properties and diversity desired. Typically, genomic DNA from organisms expressing a desired phenotype or expressing a phenotype related to the desired phenotype is utilized. Examples of such sources of genomic DNA are: genomic DNA of different species or strains of microorganisms, such as Yeast, *E.coli*, *Pseudomonads*, *Bacillus*; genomic DNA from cultured organisms originating in environments likely to encode a desired property or phenotype; genomic DNA from mixed microbial cultures or from uncultured environmental samples; genomic DNA from diversity created in the laboratory through NTG, UV mutagenesis or adaptation to certain selective conditions; and cDNA libraries of various organisms, species and strains, e.g., as indicated above, etc.

In one embodiment, the "donor DNA" and the "acceptor DNA" are pools of genomic DNA originating from the same diverse population of organisms. For example, genomic DNA from several organisms to be recombined, e.g., shuffled, is isolated. This DNA is pooled and then divided. One portion is used to construct a transposome library, the "donor DNA," while another portion is used as "acceptor DNA." In vitro transposition of the donor and acceptor pools results in the breeding of the two populations creating a combinatorial genomic library.

The source DNA is fragmented, e.g., with suitable restriction enzymes, to yield a random collection of clonable DNA fragments. These fragments are cloned between insertion sequence (IS) elements such that the genomic DNA fragments are flanked by IS elements, which under suitable conditions can transpose randomly into DNA. For example, the genomic fragments are cloned into a mini-transposon (e.g., Tn5, a shuffled mini-transposon) which contains recognition sequences (e.g., the 19-bp Tn5 transposase Mosaic End (ME) recognition sequences, inverted repeats recognized by a shuffled transposase).

The cloned library is mixed with the corresponding transposase, which binds to the recognition sequences and forms a stable complex, or transposome. Under appropriate storage conditions, e.g., Tn5 based transposomes are stable in the absence of Mg++ ions, the transposomes are stable, and can be purified and/or stored until added to a reaction mix. Genomic recombination is achieved by mixing the transposomes incorporating the donor DNA with acceptor DNA, e.g., from one or more target organisms under conditions favorable for recombination. Conditions favorable to the activity of a particular native or recombinant, e.g., shuffled, transposase can vary, and such conditions can be determined empirically to optimize recombinatorial activity of a particular transposome complex. Transposition results in the random insertion of the "mini TN library" into the acceptor DNA. The result is a library of acceptor DNA harboring integrated fragments of heterologous DNA.

In some instances, it is desirable to bias the in vitro transposition reaction with one or more nucleic acid of interest in order to create further diversity in the genomic library. This can be accomplished by spiking the reaction with transposomes including the nucleic acid of interest, such as a desired promotor, regulatory elements, e.g., terminator sequences, antiterminator sequences, Start codons, Stop codons, etc., libraries of shuffled genes, selected genes, or IS elements.

Additional diversity is introduced by performing the above process recursively. For example, a pool of recombinant nucleic acids resulting from a first in vitro transposition reaction is divided, and one portion is digested, and cloned into a mini-transposon as described above. Transposomes incorporating this new library are then prepared and used to mediate transposition, e.g., in a second portion of the recombinant nucleic acids or genomic DNA from one or more parental species or strain. This process

can be carried out for as many cycles as is desired to generate the appropriate level of diversity.

Optionally, the recombined nucleic acids are digested with suitable restriction enzymes to various sizes to facilitate their uptake and integration into host cells. These linearized fragments, or the undigested library are then delivered into suitable host cells by a variety of methods, depending on the host cell selected. For example, many microorganisms, e.g., *Bacillus Subtilis*, *Acinetobacter sp.*, *Synechocystis sp.*, *Streptococcus sp.*, etc. have natural competence mechanisms that mediate uptake of DNA molecules with high efficiency. Alternatively, the recombinant nucleic acids can be cloned into suicide vectors and introduced through standard transformation techniques such as electroporation. Suitable recipients for this approach include *E.coli*, *Saccharomyces sp.*, *Streptomyces sp.*, etc. Yet another alternative is the direct transformation, e.g., by electroporation of the recombinant nucleic acids into such host cells as yeast and other eukaryotic cells including mammalian host cells. In still another alternative, the recombinant nucleic acids are packaged into and delivered by various bacteriophages known in the art.

Following introduction of the recombinant nucleic acids into a population of host cells by any of these various means, a portion of the delivered DNA recombines with the host genome, generally by homologous recombination. This recombination results in "gene replacement" of the host DNA with the recombinant nucleic acids generated by the in vitro transposition reaction, e.g., having inserted additional material by the *in vitro* integration of the donor DNA. The resulting cell population is then screened or selected for variants having evolved toward a desired phenotype. This population is then, optionally, recombined either with itself or with other donor or acceptor DNA, and the process is repeated until the desired phenotype is achieved.

GENE IDENTIFICATION USING TRANSPOSABLE ELEMENTS

IS elements and transposons are common tools for introducing mutation in cells. These mobile genetic elements are delivered to cells using an appropriate delivery vector, tranposition is selected for and the resulting insertion mutants are screened for a phenotype of interest. Affected loci can be mapped by sequencing out from the TN into the chromosome to identify the chromosomal location. This process can be used to identify genes to be evolved, e.g., shuffled, for the improvement of desired phenotypes.

A TN harboring a drug resistance marker and origin of replication for an appropriate host organism is used to mutagenize a target organism, for example *lactobacillus*. The insertion mutants are screened for a desired phenotype, such as the ability to grow at low pH. Genomic DNA from tolerant cells is isolated and digested with a restriction enzyme not located within the TN. The digested DNA is diluted, circularized by ligation, and used to transform cells than can propagate the circularized DNA using the origin within the TN. The cloned gDNA is then sequenced to identify the affected loci. The encoded genes can then be diversified by any of the directed evolution technologies, e.g., including MolecularBreeding™, described herein, expressed in the original organism and screened for further phenotypic improvements. Alternatively, the cloned gDNA need not be sequenced, but rather can be evolved, e.g., shuffled, blindly using known sequences within the TN to tag sequences for amplification and recovery.

One such application is the identification of genomic loci that engender a desired level of gene expression. One difficulty encountered in efforts to produce improved phenotypes, is that even after optimizing a given gene contributing to the desired phenotype, significant variation can result after integration as a transgene. This is often due to differences in expression level of the optimized gene. The present invention provides vectors and methods for identifying genomic loci that result in the desired level of expression of a transgene integrated therein. For example, a target cell is co-transfected with a transiently replicating vector bearing inverted repeats, e.g., from a transposable element such as *Mariner*, a loxP site, a visible marker such as GFP and a selectable marker such as neomycin resistance. An exemplary vector is illustrated in Figure 2B. The transfected cells are exposed to neomycin and resistant cells are selected. These transfectants are then evaluated for a desired level of gene expression, e.g., GFP expression. Subsequently, a gene of interest, such as a gene optimized by shuffling, mutation or other diversity generation methods, can be integrated into the chromosomal locus by recombination at the loxP site mediated by a Cre recombinase.

GENETIC BARCODES

A further utility of using TNs, or mini-TNs, is to create tagged mutants that can be described as a composition of matter. The location of a TN within a genome of a target organism can be determined by known method, e.g., sequencing of flanking regions as described above. The TN used to create the strain can contain a predesigned

scquencc of DNA, a DNA barcode, that identifies theTN and the strain to have been created by a particular producer or manufacturer. A simple PCR reaction from the strain will amplify the sequence which can then be diagnostically sequenced to confirm its origin.

INCREASED ORGANIC ACID TOLERANCE IN *LACTOBACILLLI*

In the fermentation and bioprocess industries the optimal conditions for the organism and those for process economics do not necessarily coincide. This often poses problems of combining different phenotypes observed in various hosts into a single ideal production host, the goal being to evolvc a production host that functions under the desired conditions. In-spite-of our significant knowledge in correlating genotypes with phenotypes in well known organisms likc *E.coli*, Ycast, and *Bacillus*, it is extremely difficult to integrate multiple phenotypes into a single host using present day tools of molecular biology, classical mutagenesis, and/or metabolic engineering.

For example, a *lactobacillus* strain able to tolerate the low pH, and high concentration of organic acid required to produce high yields of lactic acid is of significant economic value. The described invention provides a method for generating such an organism. A population of *lactobacilli* each having traits desired for the industrial fermentation of lactic acid, e.g., heat tolerance, high volumetric yield, high lactic acid titer, etc., are grown and their genomic DNA (gDNA) is isolated and pooled. The gDNA is then fragmented, e.g., by limited digestion with a desired four base cutting restriction endonuclease. Fragments, typically of greater than 10 kb, are isolated and cloned within a "mini TN or IS" located on an appropriate plasmid, e.g., pTNWGS:TN5 (Figure 1B). To facilitate this cloning step, a multiple-cloning site (MCS) is positioned between the two end repeat sequences of TN5. This miniTN is flanked by the transposase gene(s) of TN5 that will catalyze, in trans, the excision and integration of the mini-TN and its contents. The plasmid pTNWGS:TN5 also contains the ColE1 origin of replication, a gene conferring positive selection in *E. coli* (such as ampicillin resistance, kanamycin resistance, chloramphenicol resistance, etc.) and in *Lactobacilli* (such as erythromycin resistance, kanamycin resistance, chloramphenicol resistance, tetracycline resistance, etc.), and a thermosensitive replicon functional in *Lacotbacillus* such as pG+.

The pTNWGS library ligation is transformed into *E.coli* (preferably deficient in restriction and modification systems). Transformants are pooled and the

plasmid DNA is isolated. The pTNWGS library is then transformed back into one or all of the starting *Lactobacilli* strains. Transformants are selected, transferred to the non-permissive temperature for pG+ and incubated to select for the loss of pTNWGS and the integration of the miniIS library into the chromosome.

The cells are then returned to the permissive temperature, and enriched for those cells having increased tolerance to low pH in the presence of organic acids. This is achieved by inoculating a turbidostat culture and continuously challenging the growing cells with medium of lower pH and increased concentrations of organic acids.

The surviving culture is separated into individual clones by plating on solid medium, and individual colonies are picked and assessed for their ability to produce high levels of lactic acid in fresh or conditioned medium. Those clones producing high levels of lactic acid are pooled, recombined (e.g., shuffled) and screened by repeating the preceding procedure. A similar protocol is employed to produce organisms that have improved performance under a variety of extreme conditions desirable for accelerated production processes, e.g., elevated temperature, high cell-density, slow growth, high end product concentration, presence of growth inhibitors or toxins, etc.

Serial fermentation for selection of improved industrial phenotypes

To facilitate the efficient and large scale improvement of industrial strains, high throughput methods requiring reduced operator involvement are preferred. One approach to increasing throughput, while reducing time and effort is by utilizing methods of selection based on the preferential survival of a subset of the population in response to selective pressures in an array of parallel continuous fermentors. A population of recombinant organisms produced by transposon diversification, e.g., shuffling, procedures is used to seed an array of parallel continuous fermentors designated f1...fx (Fig 3). The fermentors are maintained under desired selection pressures. These selection pressures need not be and most preferably are not at the level that is ultimately desired of the host. Incremental increase in selection pressures are preferred as it prevents complete wash out of the fermentors in response to the severity of the pressure. A special case arises when f1...fx are selecting a single host under incremental increases in the selection pressure (for example temperature) from one fermentor to the other.

The outlet from f11....f1n are fed to another series of parallel continuous fermentors f21....f2n where the corresponding selection pressures are increased by a small amount. A portion of the outlet streams from f21....f2n are recycled respectively to

f11....f1n. This process of recycling a cell population back to an environment of lesser intensity of the selection pressure, provides an opportunity for recuperation and expression of desired phenotype. The other portion of the outlet streams from f21..f2n are fed to a column C (WGS) which has been preconditioned for DNA exchange and uptake.

Outlet streams from f21...f2n are fed to WGS as shown in Figure 3 to foster DNA uptake between different host platforms. Conditions to enhance partial lysis of cultures to release genomic DNA, conditions to stabilize released DNA, and enhance uptake of DNA are maintained in these columns. Other variations include leaking in genomic DNA preparations from other independent experiments or sources which are believed to code for the desired phenotype.

The outlet from the WGS column is fed to another continuous fermentor f31 which is under non selective conditions to provide the opportunity to amplify the genetic diversity created in column WGS.

A portion of the outlet from f31 is distributed equally among fermentors f21...f2n to further seed them with the created diversity and thus continue with the process recursively.

The remaining part of f31 is fed to another continuous fermentor f41 which is under multiple selection pressures so as to enrich for hosts with desired multiple traits or with increased selection pressures. This fermentor is also fed with new media to dilute out strains not meeting the criteria.

Once steady state is reached and a stable population is isolated in f41, the whole process is repeated with increased selection pressures in fermentors f21..f2n. Populations isolated from f41 from the last cycle are used to seed the fermentors f21...f2n in the new cycle.

Alternatively the fermentor f41 is run as a turbidostat where all the phenotypes 1..n are gradually increased towards the desired set points in a combinatorial manner. A portion of the outlet stream from f41 is continuously fed back to the fermentors f21...f2n to further breed diversity.

As shown in Figure 3, additional genetic diversity can be introduced into the system by spiking in pools of population that have been generated or isolated by other methods independently like transposon mediated genetic diversity, conjugative libraries,

shuffled libraries and NTG/UV mutagenized pools, etc., into fermentors f11..f1x or f21..f2x..

The above protocol can be easily adapted for phenotypes for which there are no obvious selection pressure. In such cases the continuous fermentors are run under non selective conditions and their outlets are fed into various screening modules (described below in specific applications) that uses one or more criteria to enrich for desired isolates from a population. The enriched populations are fed back to the upstream fermentors and/or fed to the downstream fermentors to continue with the process. In some cases, it will be preferable to miniaturize the process on a "lab-on-a-chip" module (e.g., the LabMicrofluidic device™ high throughput screening system (HTS) by Caliper Technologies Corp., Mountain View, CA, or the HP/Agilent technologies Bioanalyzer using LabChip™ technology by Caliper Technologies Corp. *See*, also, calipertech.com) for continuous high throughput generation and selection of microbial diversity for improved phenotypes.

Application for evolving hosts with improved process phenotypes

(a) Faster Growth Rates

Improvements in growth rates of a production host has significant economic advantages. The number of batch fermentations that is typically run during a production cycle can be increased with a host that grows faster. Similarly through-put in continuous production system can be easily increased with a faster growing host. Such improvements in a production host can be achieved by the methodology described here. The selected host (s) is grown in chemostats f11...f1n (Figure 3) at different dilution rates which are proportional to their respective growth rates. The best available media is selected for this purpose and is kept fixed during the entire process. The choice of the media is often dictated by economic factors and convenience. In chemostats f21..f2n the selection pressure is further tightened by a small amount. To isolate the fastest growing host fermentor f41 is run under even higher stringency of growth rates. In cases where the primary phenotype to be conserved in the host is production of a chemical like amino acids, vitamins, neutraceuticals or a recombinant protein, the fermentor f41 is continuously monitored for productivity as the stringency on growth rate is increased. Populations that grow faster without compromising productivity are recycled to f21..f2n to continue the recursivity of the process.

Most production hosts have been evolved for expression of a primary phenotype in well defined media and process parameters. The genetic material needed to express the desired phenotype under pre-set process conditions is significantly lower than what they generally carry. Significant improvements in product yield, growth rates, and feedstock utilization can be expected by minimizing the genetic make-up (minimal genome) of these production hosts without compromising productivity of the process. For example, attempts have been made to develop identify all essential genes in the mycoplasma genome using transposon mutagenesis. The concept of minimal in this context is in terms of essential genes and not necessarily in terms of minimal physical size. Obtaining a minimal genome in terms of physical size has significant advantages as described above. Methodology described in figure 3, in combination with transposition mutagenesis, as described herein and in the references, done iteratively can be used to achieve this goal.

(b) Increased glycolysis (an example for increased feedstock uptake)

The raw material for commercial production of many biochemicals is glucose, fructose, corn starch, etc. An important economic parameter in these processes is the productivity of the process and many metabolic engineering approaches have been made to maximize this feature of a catalyst. Although these approaches are unique to the cases that they are applied to, a common feature of all these bioprocesses is that they all share a common pathway "glycolysis" by which the starting raw material glucose is processed. Ultimately the upper limit on a biotransformation rate using glucose as the feed-stock is limited by how fast a production strain can process glucose through glycolysis.

Although glycolysis is perhaps the most widely studied central metabolism pathway in microbiology, increasing the flux through this pathway (substrate uptake rate) by traditional metabolic engineering approaches have not resulted in any significant improvements. The primary reason for this is lack of significant understanding of how the components of glycolysis interact with cellular physiology and energetics under a given set of production objectives. It is also well known that flux through glycolysis increases significantly under anaerobic conditions compared to aerobic conditions in certain hosts, which suggest that the genetic components and architecture exist in microorganisms to accommodate the phenotype of increased glycolysis. The

methodology described here can be applied to evolve a host platform that expresses increased glycolytic rate under a given set of fermentation conditions.

The chosen host is grown in chemostats f11..f1n (figure 3) with the selected media and glucose as the limiting substrate. The fluoroscent glucose analog 2-NBDG is also added to these chemostats in varying concentrations from one chemostat to the other. 2-NBDG competes with D-Glucose for uptake in a competitive manner and can be monitored by microscopy or single-cell light scattering intensity. The outlet from the chemostats are fed to a cell sorter that enriches for populations that have increased uptake rates for the fluoroscent analog. A portion of this enriched populations are recycled to f21..f2x and the rest are fed to the WGS unit (Figure 3) where genomic breeding continues by one of the methods described herein. The isolation of hosts with increased glucose uptake rates will form the foundation and initial starting point for further evolution of hosts that can channel the increased glucose uptake flux to desirable products like ethanol, lactate, amino acids, isoprenoids, etc. A significant amount of research already exists for engineered hosts that efficiently channel glucose to the above described products.

Similar methodology can be easily adapted for increased uptake of other feedstocks of commercial importance.

(c) Increased TCA cycle (and pentose phosphate cycle)

The tricarboxylic acid cycle is the machinery that microorganisms use to generate energy in the form of NADH by catabolizing carbon sources into CO2. The control of flux through the TCA cycle is complicated and previous attempts to identify rate limiting steps have yielded limited success. Increasing fluxes through the TCA cycle also results in faster NADH production which is beneficial for biotransformations requiring NADH. The methodology described here can be easily adapted to evolve host platforms with increased TCA cycle flux. The flux through TCA cycle, particularly in non growing cells can be calculated from CO2 evolution rates from a chemostat. This measurement can be used to enrich for populations that have increased flux through TCA cycle for a given glucose feed rate and thus can be evolved based on the methodology suggested in Figure 3.

Similar strategies can be used to create industrial host platforms with the following attributes: increased cofactor recycling rate (cofactor engineering); decreased oxygen radicals; increased efficiency for delivering cytoplasmic molecular oxygen;

improved oxidative cytoplasm for increased efficiency of disulfide formation; increased viability in the presence of low pH, organic acids, organic solvents, desiccation, low water content, temperature (high/low), and high osmolarity.

Enrichment of viable populations under above mentioned selection pressures can be achieved using multi-staining flow cytometry as described in literature. This enrichment scheme is integrated to the outlet streams of f21..f2n and thereby enables a continuous enrichment strategy which is beneficial to evolve desired phenotypes.

In addition, the present methods can be used to produce organisms with: increased hydrophobicity (membrane properties) for improved uptake of hydrophobic compounds; improved growth properties under limiting dissolved oxygen concentrations in the fermentor; increased or sustained metabolism in the presence of high end product concentration; and organisms that utilize cheaper sources of reducing equivalents like ethanol, methanol, alkanes, etc., with high efficiency to drive biotransformations (e.g., that require reducing power).

MOLECULAR BIOLOGY

General texts which describe molecular biological techniques useful herein, including the use of vectors, promoters and many other relevant topics related to, e.g., the cloning and expression of transposable elements, transposons, insertion sequences, and their components, include Berger and Kimmel, Guide to Molecular Cloning Techniques, Methods in Enzymology volume 152 Academic Press, Inc., San Diego, CA (Berger); Sambrook et al., Molecular Cloning - A Laboratory Manual (2nd Ed.), Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, 1989 ("Sambrook") and Current Protocols in Molecular Biology, F.M. Ausubel et al., eds., Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., (supplemented through 1999) ("Ausubel")). Similarly, examples of techniques sufficient to direct persons of skill through in vitro amplification methods, including the polymerase chain reaction (PCR) the ligase chain reaction (LCR), Qβ-replicase amplification and other RNA polymerase mediated techniques (e.g., NASBA), e.g., for the production of the homologous nucleic acids of the invention are found in Berger, Sambrook, and Ausubel, as well as Mullis et al., (1987) U.S. Patent No. 4,683,202; PCR Protocols A Guide to Methods and Applications (Innis et al. eds) Academic Press Inc. San Diego, CA (1990) (Innis); Arnheim & Levinson (October 1,

1990) C&EN 36-47; The Journal Of NIH Research (1991) 3, 81-94; (Kwoh et al. (1989) Proc. Natl. Acad. Sci. USA 86, 1173; Guatelli et al. (1990) Proc. Natl. Acad. Sci. USA 87, 1874; Lomell et al. (1989) J. Clin. Chem 35, 1826; Landegren et al., (1988) Science 241, 1077-1080; Van Brunt (1990) Biotechnology 8, 291-294; Wu and Wallace, (1989) Gene 4, 560; Barringer et al. (1990) Gene 89, 117, and Sooknanan and Malek (1995) Biotechnology 13: 563-564. Improved methods of cloning in vitro amplified nucleic acids are described in Wallace et al., U.S. Pat. No. 5,426,039. Improved methods of amplifying large nucleic acids by PCR are summarized in Cheng et al. (1994) Nature 369: 684-685 and the references therein, in which PCR amplicons of up to 40kb are generated. One of skill will appreciate that essentially any RNA can be converted into a double stranded DNA suitable for restriction digestion, PCR expansion and sequencing using reverse transcriptase and a polymerase. *See*, Ausubel, Sambrook and Berger, *all supra.*

The present invention also relates to host cells and organisms which are transformed with vectors of the invention, and the production of polypeptides of the invention, e.g., transposases, exogenous DNAs incorporated into transposable elements or insertion sequences, by recombinant techniques. Host cells are genetically engineered (i.e., transformed, transduced or transfected) with the vectors of this invention, which can be, for example, a cloning vector or an expression vector. The vector can be, for example, in the form of a plasmid, a virus, a naked polynucleotide, or a conjugated polynucleotide. The vectors are introduced into cells by standard methods including electroporation (From et al. (1985) Proc. Natl. Acad. Sci. USA 82:5824, infection by viral vectors such as cauliflower mosaic virus (CaMV) (Hohn et al. (1982) Molecular Biology of Plant Tumors (Academic Press, New York) pp. 549-560; Howell, USPN 4,407,956), high velocity ballistic penetration by small particles with the nucleic acid either within the matrix of small beads or particles, or on the surface (Klein et al. (1987) Nature 327:70-73), also, especially in the case of plant cells by the use of pollen as vector (WO 85/01856), or use of *Agrobacterium tumefaciens* or *A. rhizogenes* carrying a T-DNA plasmid in which DNA fragments, e.g., including transposable elements, are cloned. The T-DNA plasmid is transmitted to plant cells upon infection by *Agrobacterium tumefaciens*, and a portion is stably integrated into the plant genome (Horsch et al. (1984) Science 233: 496-498; Fraley et al. (1983) Proc. Natl. Acad. Sci. USA 80: 4803).

The engineered host cells can be cultured in conventional nutrient media modified as appropriate for such activities as, for example, activating promoters or

selecting transformants. Where appropriate cells can be optionally cultured into transgenic organisms. For example, plant regeneration from cultured protoplasts is described in Evans et al.(1983) "Protoplast Isolation and Culture," Handbook of Plant Cell Cultures 1:124-176 (MacMillan Publishing Co., New York); Davey (1983) "Recent Developments in the Culture and Regeneration of Plant Protoplasts," Protoplasts pp. 12-29, (Birkhauser, Basel); Dale (1983) "Protoplast Culture and Plant Regeneration of Cereals and Other Recalcitrant Crops," Protoplasts pp. 31-41, (Birkhauser, Basel); Binding (1985) "Regeneration of Plants," Plant Protoplasts pp. 21-73, (CRC Press, Boca Raton).

The present invention also relates to the production of transgenic organisms, which can be bacteria, yeast, fungi, or plants. A thorough discussion of techniques relevant to bacteria, unicellular eukaryotes and cell culture can be found in references enumerated above and are briefly outlined as follows. Several well-known methods of introducing target nucleic acids into bacterial cells are available, any of which can be used in the present invention. These include: fusion of the recipient cells with bacterial protoplasts containing the DNA, electroporation, projectile bombardment, and infection with viral vectors (discussed further, below), etc. Bacterial cells can be used to amplify the number of plasmids containing DNA constructs of this invention. The bacteria are grown to log phase and the plasmids within the bacteria can be isolated by a variety of methods known in the art (*see*, for instance, Sambrook). In addition, a plethora of kits are commercially available for the purification of plasmids from bacteria. For their proper use, follow the manufacturer's instructions (see, for example, EasyPrep™, FlexiPrep™, both from Pharmacia Biotech; StrataClean™, from Stratagene; and, QIAprep™ from Qiagen). The isolated and purified plasmids are then further manipulated to produce other plasmids, used to transfect plant cells or incorporated into *Agrobacterium tumefaciens* related vectors to infect plants. Typical vectors contain transcription and translation terminators, transcription and translation initiation sequences, and promoters useful for regulation of the expression of the particular target nucleic acid. The vectors optionally comprise generic expression cassettes containing at least one independent terminator sequence, sequences permitting replication of the cassette in eukaryotes, or prokaryotes, or both, (e.g., shuttle vectors) and selection markers for both prokaryotic and eukaryotic systems. Vectors are suitable for replication and integration in prokaryotes, eukaryotes, or preferably both. *See*, Giliman & Smith

(1979) Gene 8:81; Roberts et al. (1987) Nature 328:731; Schneider et al. (1995) Protein Expr. Purif. 6435:10; Ausubel, Sambrook, Berger (*all supra*). A catalogue of Bacteria and Bacteriophages useful for cloning is provided, e.g., by the ATCC, e.g., The ATCC Catalogue of Bacteria and Bacteriophage (1992) Gherna et al. (eds) published by the ATCC. Additional basic procedures for sequencing, cloning and other aspects of molecular biology and underlying theoretical considerations are also found in Watson et al. (1992) Recombinant DNA (Second Edition) Scientific American Books, NY.

While the foregoing invention has been described in some detail for purposes of clarity and understanding, it will be clear to one skilled in the art from a reading of this disclosure that various changes in form and detail can be made without departing from the true scope of the invention. For example, all the techniques, methods, compositions, apparatus and systems described above may be used in various combinations. All publications, patents, patent applications, or other documents cited in this application are incorporated by reference in their entirety for all purposes to the same extent as if each individual publication, patent, patent application, or other document were individually indicated to be incorporated by reference for all purposes.

WHAT IS CLAIMED IS:

1. A method for producing one or more transposable element component with a desired property, the method comprising:

i) providing a population of polynucleotide segments comprising at least one transposable element or subportion of a transposable element;

ii) recombining the polynucleotide segments one or more times, thereby producing a library of recombinant transposable element components;

iii) identifying at least one recombinant transposable element component with a desired property;

iv) optionally repeating steps (i) through (iii) at least one additional time.

2. The method of claim 1, further comprising recovering the transposable element or a subportion thereof by a polymerase chain reaction (PCR), ligase chain reaction (LCR), Qβ-replicase amplification, NASBA or cloning.

3. The method of claim 1, comprising providing a population of polynucleotide segments comprising at least one component of a transposon or insertion sequence (IS) element or a subportion of a component of a transposon or IS element.

4. The method of claim 3, wherein the at least one component of a transposable element comprises an inverted repeat or a transposase of a transposon or an IS element.

5. The method of claim 3, wherein the transposon or IS element comprises a mini-transposon or a mini-IS element.

6. The method of claim 1, wherein at least one polynucleotide segment comprises a transposable element or a subportion of a transposable element of a bacterium, a fungus, a plant or an animal.

7. The method of claim 1, wherein the transposable element comprises a Class I or a Class II transposable element.

8. The method of claim 7, wherein the Class I transposable element comprises a retrotransposon, a retroposon or a SINE-like element.

9. The method of claim 8, wherein the Class I transposable element comprises a *Ty-1* family transposon, a *Copia* family transposon, or a *gypsy* family transposon.

10. The method of claim 7, wherein the Class II transposable element comprises a *Fotl/Pogo* family transposon, a *Tcl/Mariner* family transposon,

11. The method of claim 7, wherein the transposable element is selected from the group consisting of TN3, TN5, TN10, TN917, ISS1, TN5990, Ty1, Ty2, Ty3, and mariner.

12. The method of claim 1, comprising recombining the polynucleotide segments in vitro, in vivo, or in silico.

13. The method of claim 1, wherein the desired property is selected from one or more of altered specificity of integration, host adaptation, altered cofactor specificity, increased or decreased recombinase activity, increased or decreased transposase activity, increased or decreased recombinase specificity, increased or decreased transposase specificity, increased or decreased size of exogenous DNA transposed, increased or decreased copy number, increased or decreased efficiency of transposition, increased or decreased preference for episomal targeting, increased or decreased preference for chromosomal targeting, increased efficiency of integration into non-supercoiled DNA, and increased efficiency of in vitro transposition.

14. The method of claim 1, wherein the identifying of step (iii) comprises screening or selecting at least one transposable element with a desired property.

15. The method of claim 14, comprising identifying at least one transposable element that mediates transposition in vitro with greater efficiency when compared to a parental transposable element, the method comprising:
providing a plurality of in vitro transposition reactions, which in vitro transposition reactions comprise:
(a) a transposase;

(b) a donor polynucleotide comprising at least one inverted repeat; and

(c) a target polynucleotide

incubating the plurality of in vitro transposition reactions under conditions permissive for in vitro transposition; and

identifying at least one in vitro transposition reaction that occurs with greater efficiency than an in vitro transposition reaction mediated by a parental transposable element.

16. The method of claim 15, comprising providing in vitro transposition reactions comprising transposomes, which transposomes comprise the transposase and the donor polynucleotide.

17. The method of claim 14, comprising identifying at least one transposable element that transposes with increased efficiency in a specified host cell when compared with a wild type transposable element, the method comprising:

(a) introducing a plurality of transposable elements, which transposable elements differ by at least one nucleotide, into a population of host cells;

(b) selecting at least one host cell that has integrated the transposable element into a chromosome or episome.

18. The method of claim 17, the transposable element comprising in the direction of transcription (a) a polynucleotide comprising a transcription regulatory sequence; (b) a 5' splice donor site; (c) a first inverted repeat; (d) a 3' splice acceptor site; (e) a polynucleotide encoding a transposase; (f) a polynucleotide encoding a selectable marker; and (g) a second inverted repeat.

19. The method of claim 18, which transposable element transiently expresses the transposase.

20. The method of claim 17, comprising selecting at least one host cell that expresses a sufficient level of a selectable marker encoded by the transposable element.

21. The method of claim 17, further comprising recovering the transposable element.

22. The method of claim 17, the population of host cells comprising mammalian cells.

23. The method of claim 17, wherein the transposable element comprises a *Mariner* transposase, and wherein the inverted repeats comprise *Mariner* inverted repeats.

24. The method of claim 23, wherein the *Mariner* transposase comprises a Himar1 transposase.

25. The method of claim 17, wherein the selectable marker comprises drug resistance.

26. The method of claim 25, wherein the antibiotic resistance is selected from among neomycin resistance, kanamycin resistance.

27. The method of claim 1, wherein the transposable element comprises a recombinant vector.

28. The method of claim 27, wherein the recombinant vector is a delivery vector comprising (a) an origin of replication active in at least one cloning host; (b) a conditional origin of replication active in at least one target cell; (c) at least one screenable or selectable marker; (d) a mini-transposon comprising a first inverted repeat and a second inverted repeat, which inverted repeats flank a multicloning site (MCS); and (e) a transposase operably linked to a promoter active in at least one target cell.

29. The method of claim 28, wherein the transposase is in close proximity to at least one end of the minitransposon.

30. The recombinant delivery vector of claim 28.

31. The recombinant delivery vector of claim 30, the origin of replication (a) comprising an origin of replication selected from among a ColE1 origin, a pACYC origin, a p15A origin, an RK4 origin, an RK6 origin, a pCM595 origin, a pSa origin, a pUB110 origin, a pE194 origin, a pG+ origin, a 2 micron circle origin, and an artificial chromosome origin.

32. The recombinant delivery vector of claim 30, the conditional origin of replication (b) comprising a temperature sensitive origin of replication selected from among a gram negative origin, a pSA3 origin, a pE194tm origin, and a pG+tm origin.

33. The recombinant delivery vector of claim 30, the at least one selectable or screenable marker (c) comprising antibiotic resistance, conferred prototrophy, or toxicity resistance.

34. The recombinant delivery vector of claim 31, wherein the antibiotic resistance marker comprises kanamycin resistance, ampicillin resistance, macolide-lincosamine-streptogramin (MLS) resistance, apramycin resistance, spiramycin resistance, hygromycin resistance, chloramphenicol resistance, or tetracycline resistance.

35. The recombinant delivery vector of claim 30, wherein the mini-transposon (d) is derived from a transposon or insertion sequence of table 1.

36. The recombinant delivery vector of claim 30, the transposase (e) comprising a naturally occuring transposase or a transposase derived by one or more directed evolution method.

37. The recombinant delivery vector of claim 36, wherein the promoter is selected from an endogenous promoter of a target cell.

38. The recombinant delivery vector of claim 30, comprising in the order of transcription: a polynucleotide encoding a transposase operably linked to a promoter functional in a target cell, a mini IS element, which mini-IS element comprises a first IS inverted repeat and a second IS inverted repeat, which first and second IS inverted repeats flank a multicloning site, a first origin of replication functional in cloning host, a first selectable marker, a second selectable marker, and a second origin of replication, which origin of replication is temperature sensitive.

39. The recombinant delivery vector of claim 38, wherein the transposase comprises: a transposon or IS element int encoding sequence and a transposon or IS element xis encoding sequence.

40. The recombinant delivery vector of claim 38, wherein the first or second selectable marker comprises a drug resistance marker selected from: ampicillin resistance, kanamycin resistance, chloramphenicol resistance, neomycin resistance, tetracycline resistance, erythromycin resistance and G418 resistance.

41. The recombinant delivery vector of claim 38, wherein the first and second selectable markers comprise two alternative markers selected from: ampicillin resistance, kanamycin resistance, chloramphenicol resistance, neomycin resistance, tetracycline resistance, erythromycin resistance and G418 resistance.

42. The recombinant delivery vector of claim 38, wherein the second origin of replication comprises a thermosensitive replicon of pG+.

43. The recombinant delivery vector of claim 38, wherein the vector comprises a pTNWGS vector.

44. A transposable element with a desired property produced by the method of claim 1.

45. The transposable element of claim 44, wherein the desired property is selected from one or more of altered specificity of integration, host adaptation, increased or decreased recombinase activity, increased or decreased transposase activity, increased or decreased recombinase specificity, increased or decreased transposase specificity, increased or decreased size of exogenous DNA transposed, increased or decreased copy number, increased or decreased efficiency of transposition, increased or decreased preference for episomal targeting, increased or decreased preference for chromosomal targeting, increased efficiency of integration into non-supercoiled DNA, and increased efficiency of in vitro transposition.

46. The transposable element of claim 44, which transposable element catalyzes in vitro transposition more efficiently than a parental transposable element.

47. The transposable element of claim 44, which transposable element integrates into a specified host cell with increased efficiency when compared to a wild type transposable element.

48. A component of a transposable element with a desired property produced by the method of claim 1.

49. The transposable element component of claim 48, wherein the component comprises a transposase, a recombinase or an integrase.

50. The transposable element component of claim 49, comprising a transposase that catalyzes in vitro transposition more efficiently than a parental transposase.

51. The transposable element component of claim 48, wherein the component comprises an inverted repeat.

52. A method for producing a transposase that efficiently catalyzes in vitro transposition, the method comprising:

i) providing a population of polynucleotide segments encoding at least one transposase or subportion of a transposase;

ii) recombining the polynucleotide segments one or more times, thereby producing a library of recombinant polynucleotides encoding variant transposases;

iii) identifying at least one recombinant polynucleotide encoding a transposase that efficiently catalyzes in vitro transposition.

53. The method of claim 52, comprising identifying the at least one recombinant polynucleotide encoding a transposase that efficiently catalyzes in vitro transposition by:

a) providing a plurality of in vitro transposition reactions, which in vitro transposition reactions comprise a transposase encoded by the recombinant polynucleotide, a donor polynucleotide comprising at least one inverted repeat, and a target polynucleotide;

b) incubating the plurality of in vitro transposition reactions under conditions permissive for in vitro transposition; and

c) identifying at least one in vitro transposition reaction that occurs with greater efficiency than an in vitro transposition reaction mediated by a parental transposase.

54. A transposase produced by the method of claim 52.

55. The transposasc of claim 54, wherein the transposase is selected from among transposases derived by a directed evolution process from at least one transposase of TN5, TN10, TN917, ISS1, TN5990, Ty1, Ty2, Ty3, or mariner.

56. A reaction mix or a cell comprising the transposase of claim 54.

57. A method for generating diversity in a population of nucleic acids, the method comprising: contacting at least one recombinant transposable element or recombinant transposable element component, and a plurality of subject nucleic acids under conditions permissive for transposition.

58. The method of claim 57, wherein the recombinant transposable element or recombinant transposable element component is produced by one or more diversity generating procedure.

59. The method of claim 57, wherein the recombinant transposable element or recombinant transposable element component is produced by recursive recombination.

60. The method of claim 57, further comprising identifying at least one altered subject nucleic acid.

61. The method of claim 57, comprising contacting the recombinant transposable element or recombinant transposable element component and the subject nucleic acids in vivo.

62. The method of claim 61, wherein the recombinant transposable element component comprises a recombinant transposase.

63. The method of claim 62, comprising introducing a transposome, which transposome comprises the recombinant transposase bound to a donor nucleic acid, which donor nucleic acid comprises sequences recognized by the recombinant transposase, into a cell, thereby contacting the recombinant transposable element component and the subject nucleic acids.

64. The method of claim 63, comprising introducing the transposome into the cell by electroporation.

65. The method of claim 57, comprising contacting the transposable element or transposable element component and the subject nucleic acids in vitro.

66. The method of claim 65, wherein the recombinant transposable element component comprises a recombinant transposase.

67. The method of claim 65, comprising contacting the subject nucleic acids with a transposome, which transposome comprises the shuffled transposase bound to a donor nucleic acid, which donor nucleic acid comprises sequences recognized by the shuffled transposase, in an acellular reaction mix.

68. A method for generating diversity in a population of nucleic acids, the method comprising:

i) providing a plurality of transposomes, which transposomes comprise a library of donor nucleic acids, and a population of acceptor nucleic acids in vitro;

ii) recombining the donor nucleic acids and the acceptor nucleic acids to produce a library of recombinant nucleic acids.

69. The method of claim 68, comprising recombining the donor nucleic acids and the acceptor nucleic acids in the presence of magnesium ions.

70. The method of claim 68, comprising providing the transposome by combining a plurality of donor nucleic acid molecules, which donor nucleic acid molecules comprise transposable element recognition sequences and a plurality of transposase molecules, which transposase molecules bind the transposable element recognition sequences.

71. The method of claim 70, wherein the donor nucleic acids comprising transposable element recognition sequences are produced by cloning genomic DNA fragments into a mini-transposon or mini-insertion sequence.

72. The method of claim 71, wherein the genomic DNA fragments are restriction enzyme fragments.

73. The method of claim 71, wherein the mini-transposon comprises a Tn5 mini-transposon.

74. The method of claim 71, wherein the mini-transposon comprises a mariner transposon.

75. The method of claim 68, wherein one or more of the donor or acceptor nucleic acids are derived from a plurality of organisms.

76. The method of claim 68, further providing at least one population of additional nucleic acids.

77. The method of claim 76, the population of additional nucleic acids comprising one or more of a promoter, a regulatory element, a terminator sequence, an antiterminator sequence, a sequence comprising a start codon, a sequence comprising a stop codon, a library of recombinant genes, a gene of interest, or an IS element.

78. The method of claim 68, further comprising repeating the recombination of steps i) and ii) by providing transposomes comprising the library of recombinant nucleic acids or a subportion thereof.

79. The method of claim 68, further comprising, introducing the library of recombinant nucleic acids or a subportion thereof into a population of cells and identifying at least one cell with a desired property.

80. The method of claim 79, comprising introducing the library of recombinant nucleic acids or a subportion thereof into the population of cells by a delivery method comprising natural competence, conjugation, transformation, electroporation, or infection with bacteriophage.

81. A method for identifying a chromosomal locus, which chromosomal locus exhibits a desired level of gene expression, the method comprising:

i) transfecting a plurality of host cells expressing a transposase with a vector comprising, in the direction of transcription: (a) a first inverted repeat; (b) a promoter; (c) a site specific recombinase recognition site; (d) a polynucleotide

encoding a first screenable or selectable marker; (e) a polynucleotide encoding a second screenable or selectable marker; and (f) a second inverted repeat;

ii) identifying at least one host cell that expresses a sufficient level of at least one selectable marker, which selectable marker is encoded by the first or second visible or selectable marker, to survive selection, thereby identifying at least one host cell that has integrated the vector into a chromosome; and

iii) identifying at least one host cell expressing at least one screenable or selectable marker at a desired level, thereby identifying a chromosomal locus exhibiting a desired level of gene expression.

82. The method of claim 81, wherein the vector further comprises a polynucleotide encoding the transposase operably linked to a promoter active in the host cells.

83. The method of claim 81, further comprising integrating a polynucleotide sequence of interest into the identified chromosomal locus to generate at least one integrant.

84. The method of claim 82, further comprising identifying at least one integrant with a desired level of expression.

85. The method of claim 81, wherein the inverted repeats comprise transposable element inverted repeats.

86. The method of claim 85, wherein the inverted repeats comprise *Mariner* inverted repeats.

87. The method of claim 81, wherein the site specific recombinase recognition site comprises a loxP site.

88. The method of claim 81, wherein the promoter comprises a cytomegalovirus (CMV) promoter.

89. The method of claim 81, wherein the first or second screenable or selectable marker is a selectable marker selected from among: antibiotic resistance, herbicide resistance, neomycin resistance, kanamycin resistance.

90. The method of claim 81, whcrein thc fist or second screenable or selectable marker is a visible marker selected from among: green fluorescent protein (GFP), luciferase, β-galactosidase, β-glucuronidase, alkaline phosphatase.

91. The method of claim 81, wherein the first screenable or selectable marker comprises a visible marker and the second screenable or selectable marker comprises a selectable marker.

92. The method of claim 91, wherein the visible marker is GFP and the selectable marker is neomycin resistance.

93. The method of claim 81, the plurality of cells comprising bacterial, fungal, animal or plant cells.

94. The method of claim 81, wherein the transposase is encoded by a chromosomal sequence.

95. The method of claim 81, wherein the transposase is encoded by a polynucleotide comprising an additional vector.

96. The method of claim 95, wherein the additional vector comprises an episomal vector.

97. The method of claim 95, wherein the vector comprises a chromosomally integrated vector.

98. The method of claim 95, comprising expressing the transposase transiently.

99. The method of claim 98, comprising expressing the transposase inducibly.

100. The method of claim 81, comprising expressing a *Mariner* transposase.

101. The method of claim 100, wherein the transposase comprises an artificially evolved transposase, which artificially evolved transposase has at least one

property which differs from a parental transposase from which it is derived by directed evolution.

102. The method of claim 101, wherein the at least one property which differs from the parental transposase is selected from among: sequence specificity, activity level, species selectivity, allostery, and control.

103. A vector comprising (a) a first inverted repeat; (b) a promoter; (c) a site specific recombinase recognition site; (d) a polynucleotide encoding a first screenable or selectable marker; (e) a polynucleotide encoding a second screenable or selectable marker; and (f) a second inverted repeat.

104. The vector of claim 103, wherein the inverted repeats comprise transposable element inverted repeats.

105. The vector of claim 104, wherein the inverted repeats comprise *Mariner* inverted repeats.

106. The vector of claim 103, wherein the site specific recombinase recognition site is a loxP site.

107. The vector of claim 103, wherein the promoter comprises a cytomegalovirus (CMV) promoter.

108. The vector of claim 103, wherein the first or second screenable or selectable marker is a selectable marker selected from among: antibiotic resistance, herbicide resistance, neomycin resistance, kanamycin resistance.

109. The vector of claim 103, wherein the first or second screenable or selectable marker is a visible marker selected from among: green fluorescent protein (GFP), luciferase, β-galactosidase, β-glucuronidase, alkaline phosphatase.

110. The vector of claim 103, wherein the first screenable or selectable marker comprises a visible marker and the second visible or selectable marker comprises a selectable marker.

111. The vector of claim 103, wherein the visible marker is GFP and the selectable marker is neomycin resistance.

112. A vector comprising in the direction of transcription: (a) a polynucleotide comprising a transcription regulatory sequence; (b) a 5' splice donor site; (c) a first inverted repeat; (d) a 3' splice acceptor site; (e) a polynucleotide encoding a transposase; (f) a polynucleotide encoding a selectable marker; and (g) a second inverted repeat.

113. The vector of claim 103, wherein the first and second inverted repeat comprises *Mariner* inverted repeats.

114. The vector of claim 103, wherein the transposase comprises a *Mariner* transposase.

115. The vector of claim 103, wherein the first and second inverted repeats comprise *Mariner* inverted repeats, and the transposase comprises a *Mariner* transposase.

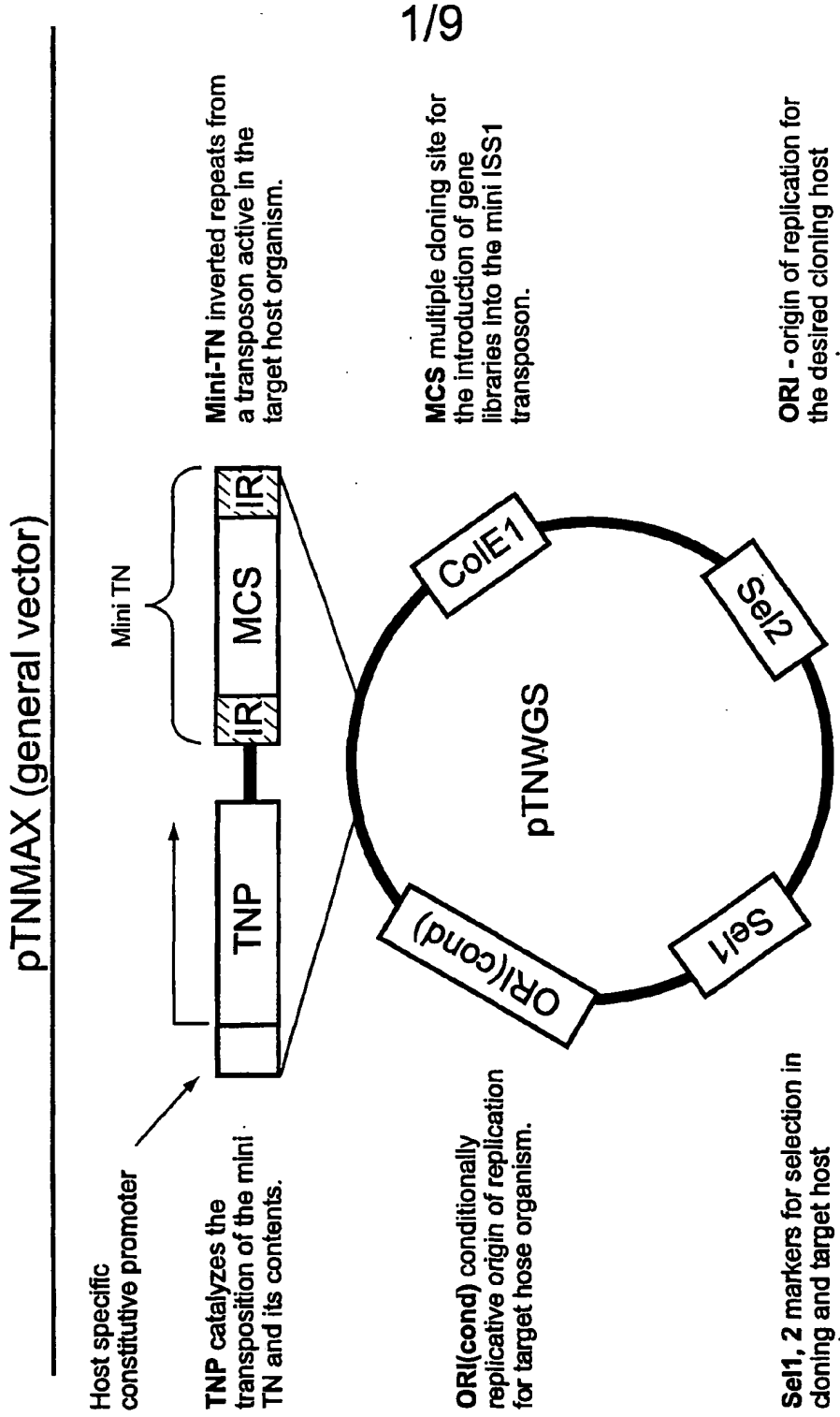

pTNMAX (general vector)
Host specific constitutive promoter
TNP
Mini TN
IR
MCS
IR
TNP catalyzes the transposition of the mini TN and its contents.
Mini-TN inverted repeats from a transposon active in the target host organism.
MCS multiple cloning site for the introduction of gene libraries into the mini ISS1 transposon.
ORI - origin of replication for the desired cloning host organism.
pTNWGS
ColE1
Sel2
Sel1
ORI(cond)
ORI(cond) conditionally replicative origin of replication for target hose organism.
Sel1, 2 markers for selection in cloning and target host organism.

Fig. 1A

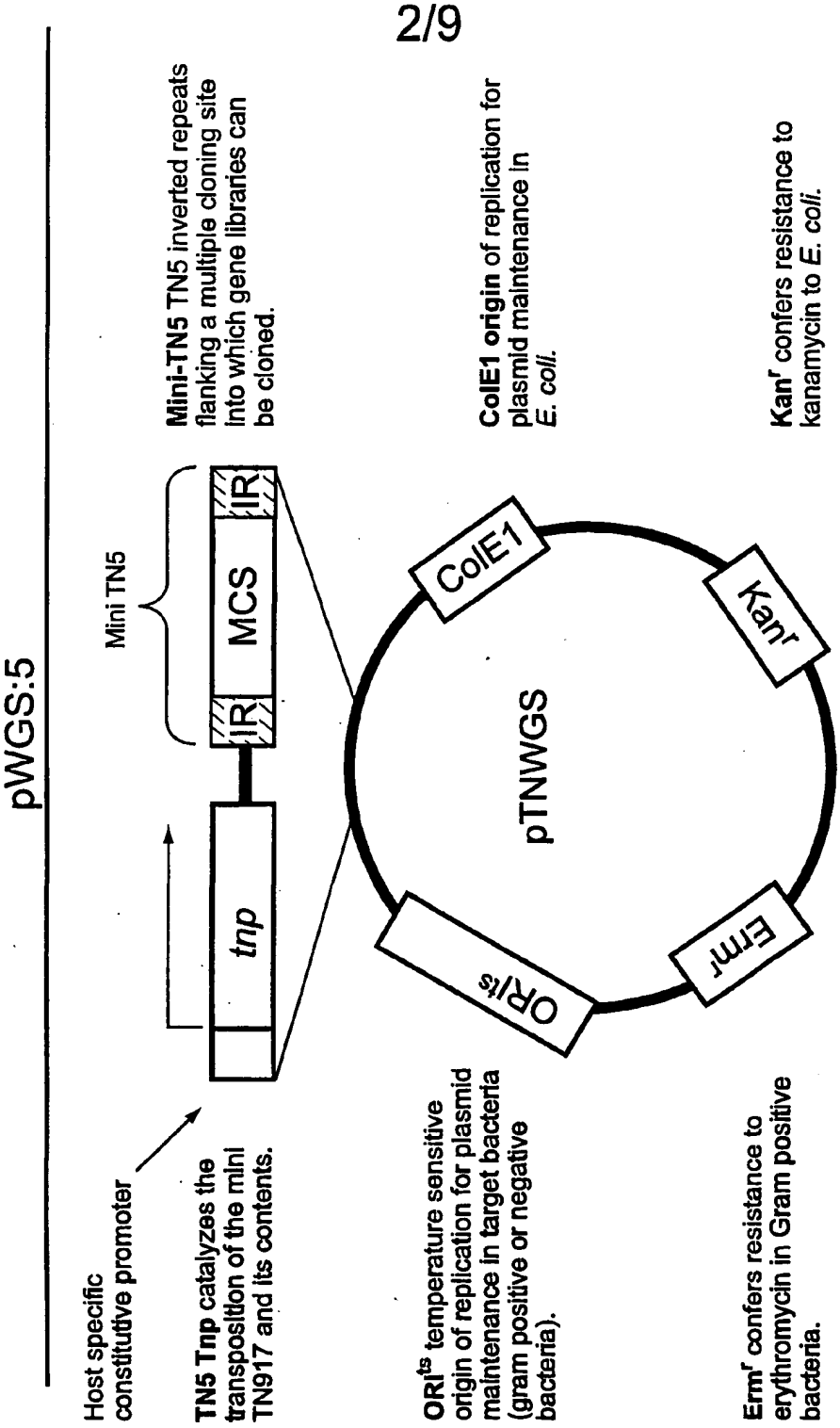

pWGS:5
Host specific constitutive promoter
TN5 Tnp catalyzes the transposition of the mini TN917 and its contents.
tnp
Mini TN5
IR
MCS
IR
Mini-TN5 TN5 inverted repeats flanking a multiple cloning site into which gene libraries can be cloned.
pTNWGS
ColE1
ColE1 origin of replication for plasmid maintenance in E. coli.
Kanr
Kanr confers resistance to kanamycin to E. coli.
ORIts
ORIts temperature sensitive origin of replication for plasmid maintenance in target bacteria (gram positive or negative bacteria).
Ermr
Ermr confers resistance to erythromycin in Gram positive bacteria.

Fig. 1B

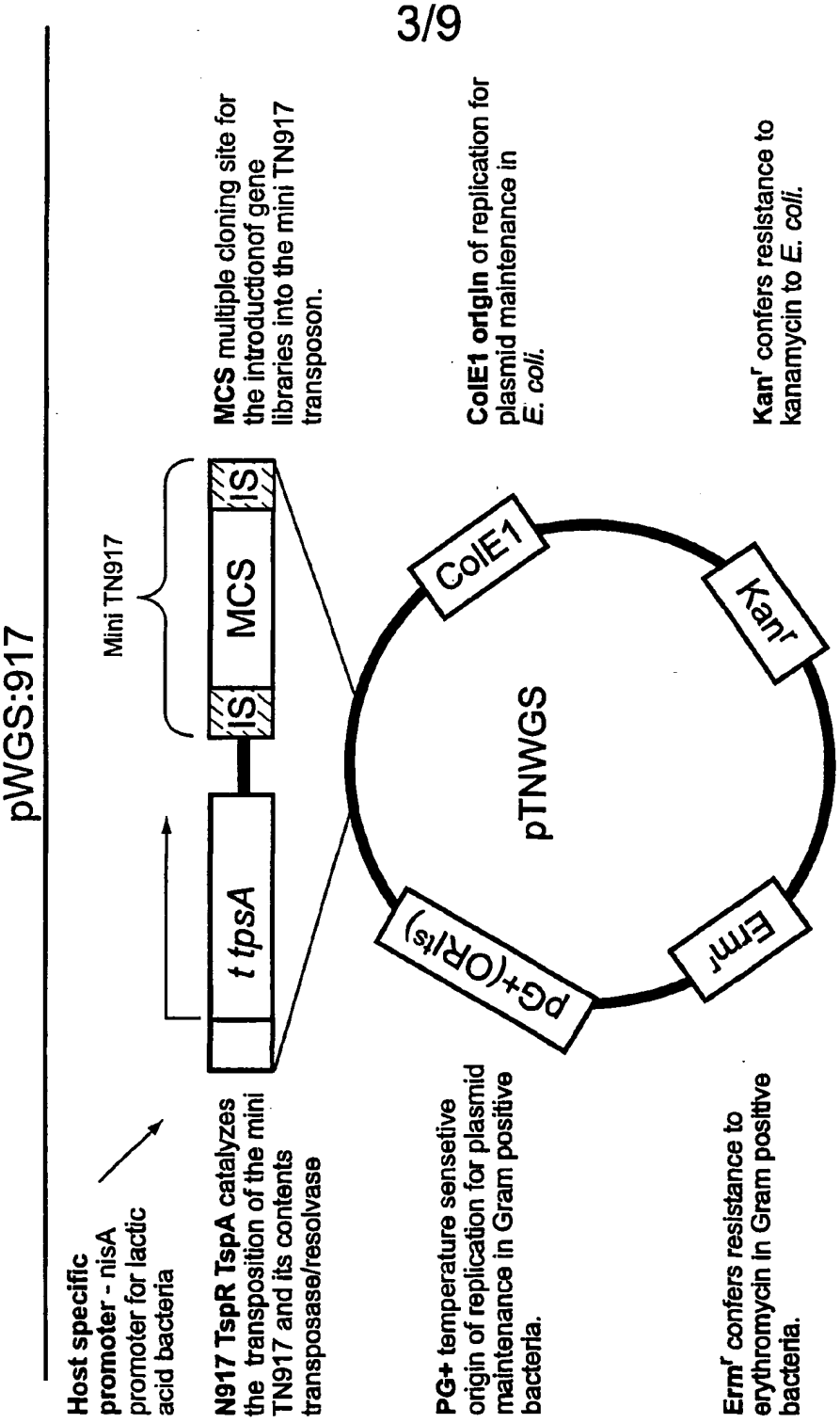

pWGS:917
Host specific promoter - nisA promoter for lactic acid bacteria
N917 TspR TspA catalyzes the transposition of the mini TN917 and its contents transposase/resolvase
PG+ temperature sensetive origin of replication for plasmid maintenance in Gram positive bacteria.
Erm^r confers resistance to erythromycin in Gram positive bacteria.
Kan^r confers resistance to kanamycin to E. coli.
ColE1 origin of replication for plasmid maintenance in E. coli.
MCS multiple cloning site for the introductionof gene libraries into the mini TN917 transposon.
Mini TN917
t tpsA
IS
MCS
IS
IS
pTNWGS
ColE1
Kan^r
Erm^r
PG+(ORI^ts)

Fig. 1C

Efficient integration into mammalian cells using evolved *Mariner* transposons

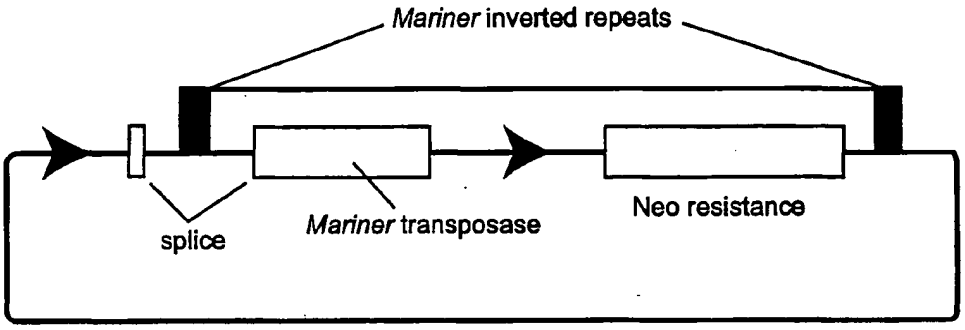


Fig. 2A

Mariner transposons for inserting loxP sites at loci with desirable expression properties

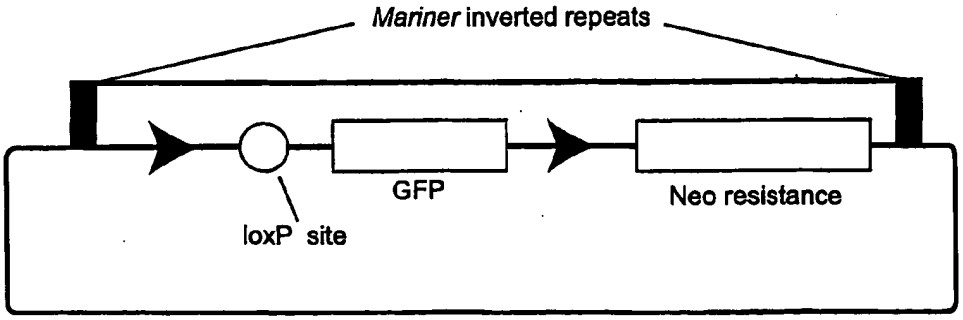


Fig. 2B

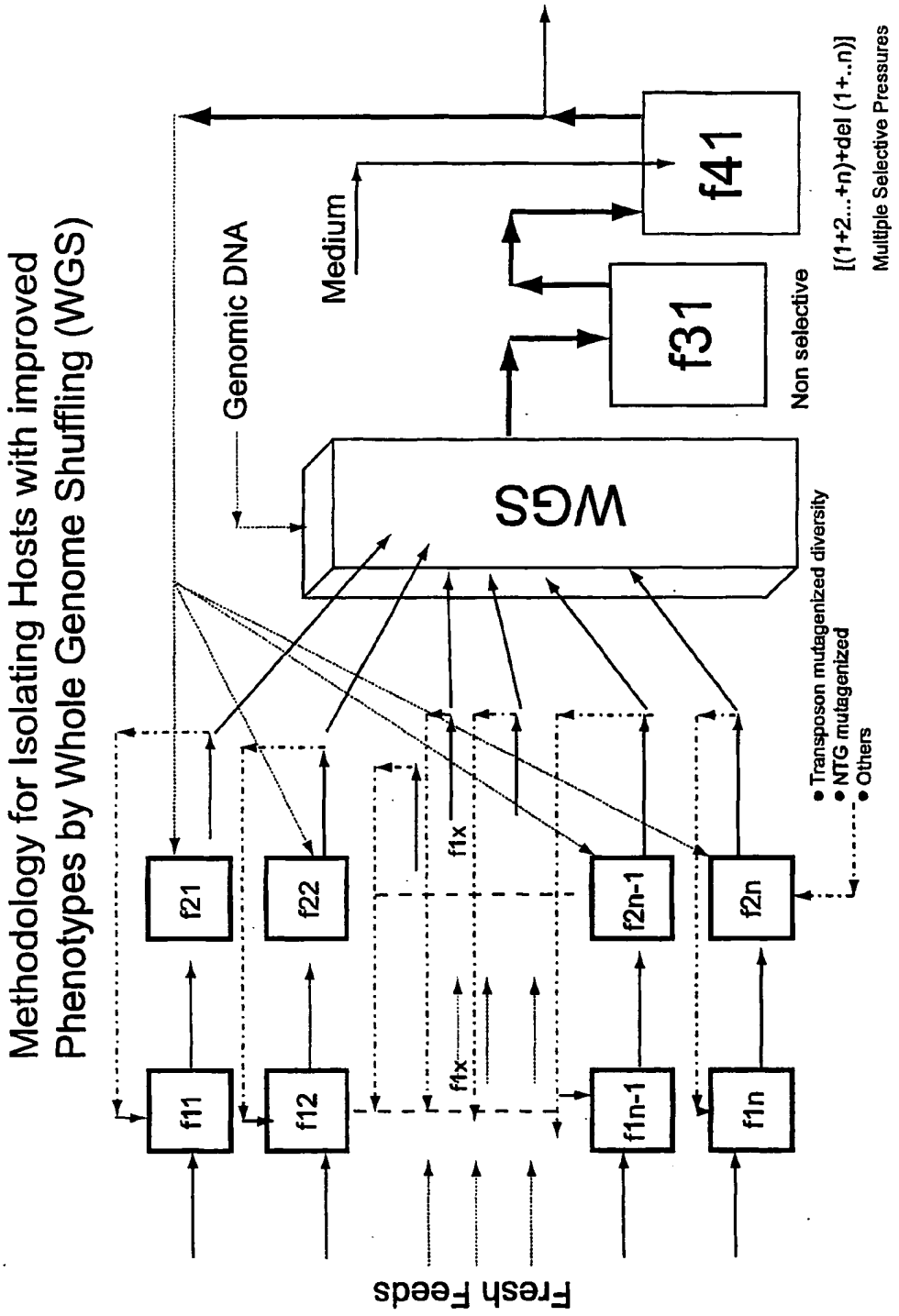

Methodology for Isolating Hosts with improved Phenotypes by Whole Genome Shuffling (WGS)
Fresh Feeds
f11
f12
f1x
f1n-1
f1n
f21
f22
f2n-1
f2n
WGS
Genomic DNA
Medium
f31
f41
Non selective
[(1+2...+n)+del (1+..n)]
Multiple Selective Pressures
• Transposon mutagenized diversity
• NTG mutagenized
• Others

Fig. 3

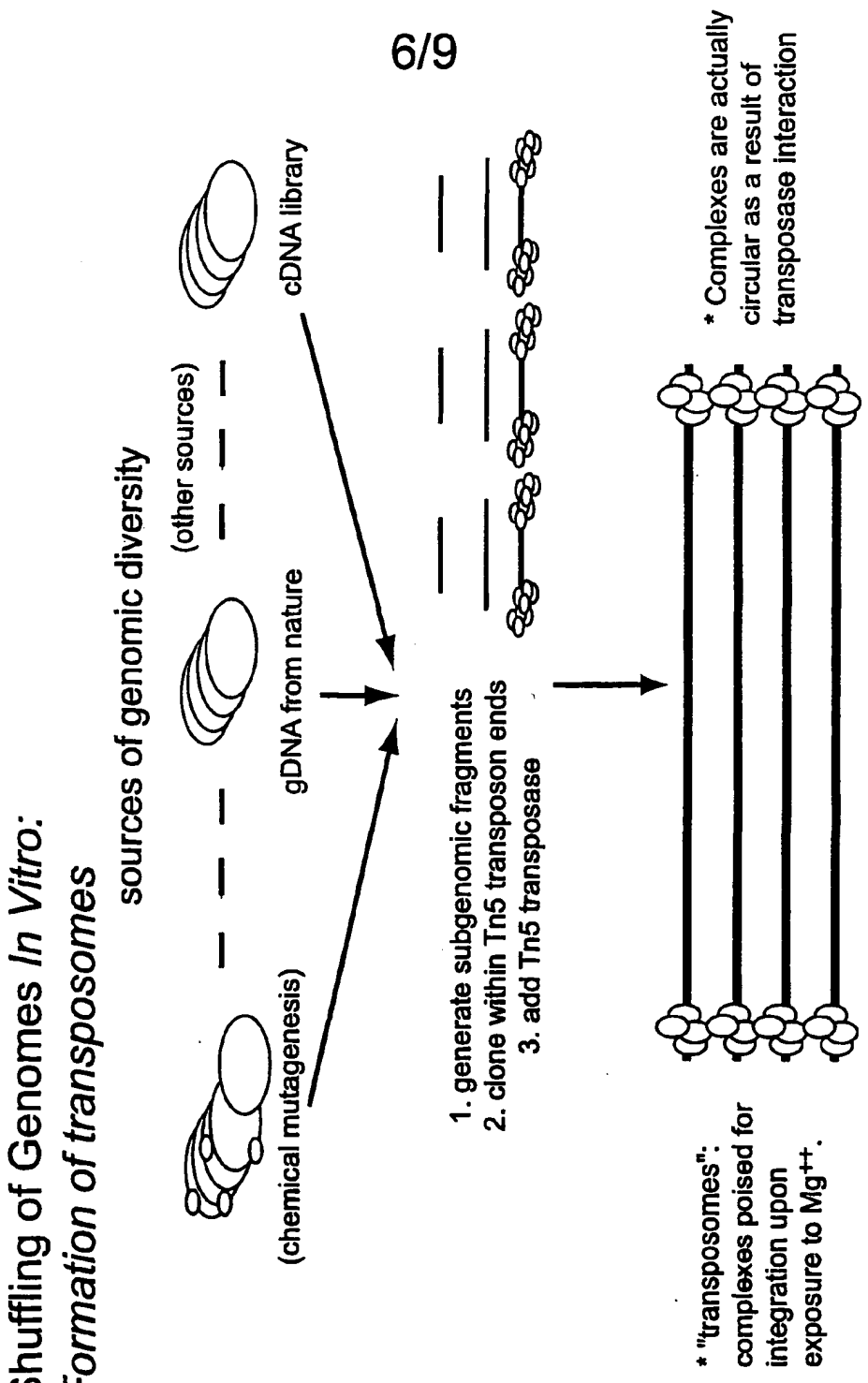

Shuffling of Genomes *In Vitro*:
Formation of transposomes
sources of genomic diversity
(chemical mutagenesis)
gDNA from nature
(other sources)
cDNA library
1. generate subgenomic fragments
2. clone within Tn5 transposon ends
3. add Tn5 transposase
* "transposomes": complexes poised for integration upon exposure to Mg^{++}.
* Complexes are actually circular as a result of transposase interaction

Fig. 4A

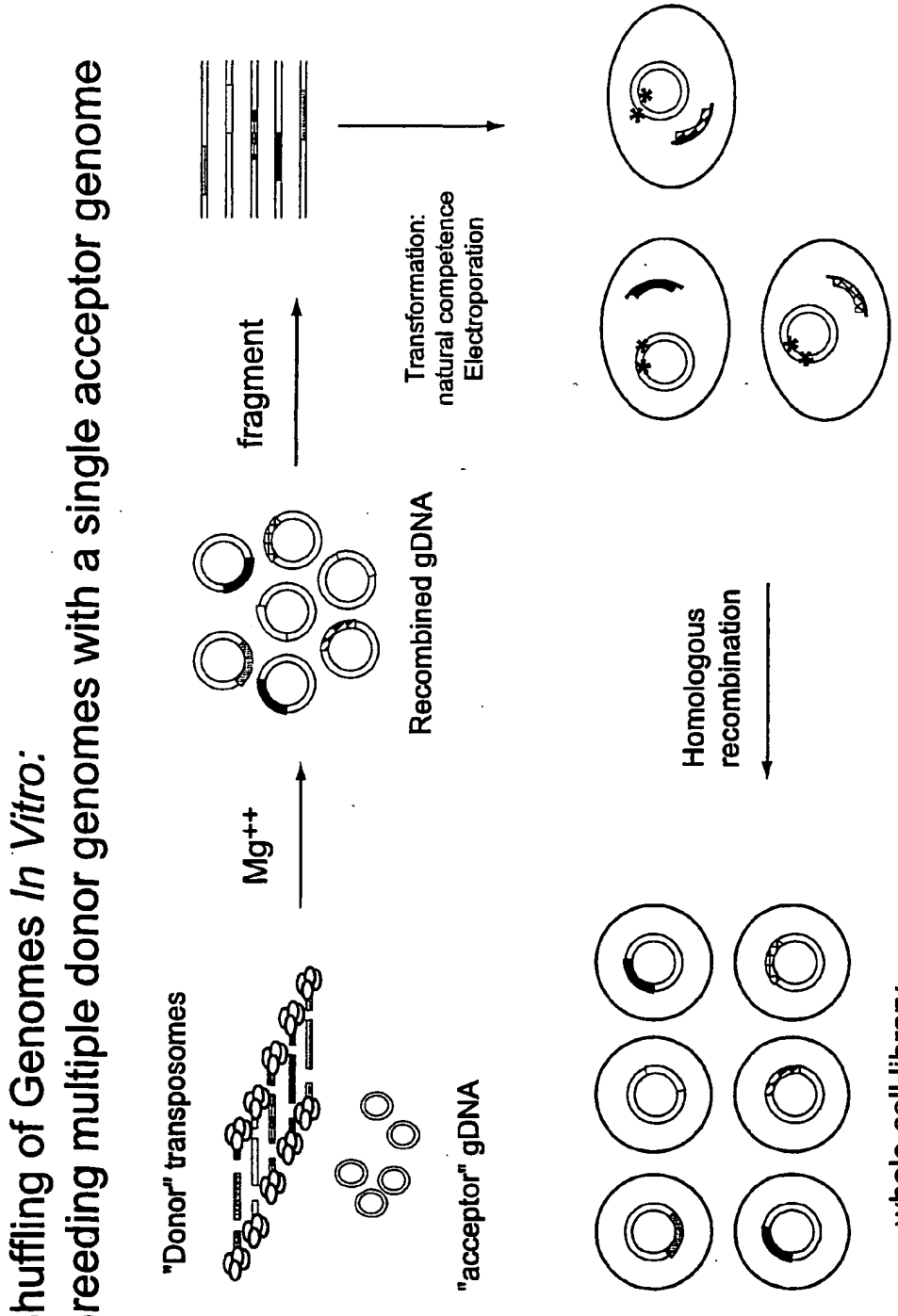

Shuffling of Genomes *In Vitro*:
Breeding multiple donor genomes with a single acceptor genome
"Donor" transposomes
"acceptor" gDNA
Mg++
Recombined gDNA
fragment
Transformation:
natural competence
Electroporation
Homologous
recombination
whole cell library

Fig. 4B

Shuffling of Genomes *In Vitro:*
Breeding multiple donor genomes with multiple acceptor genome

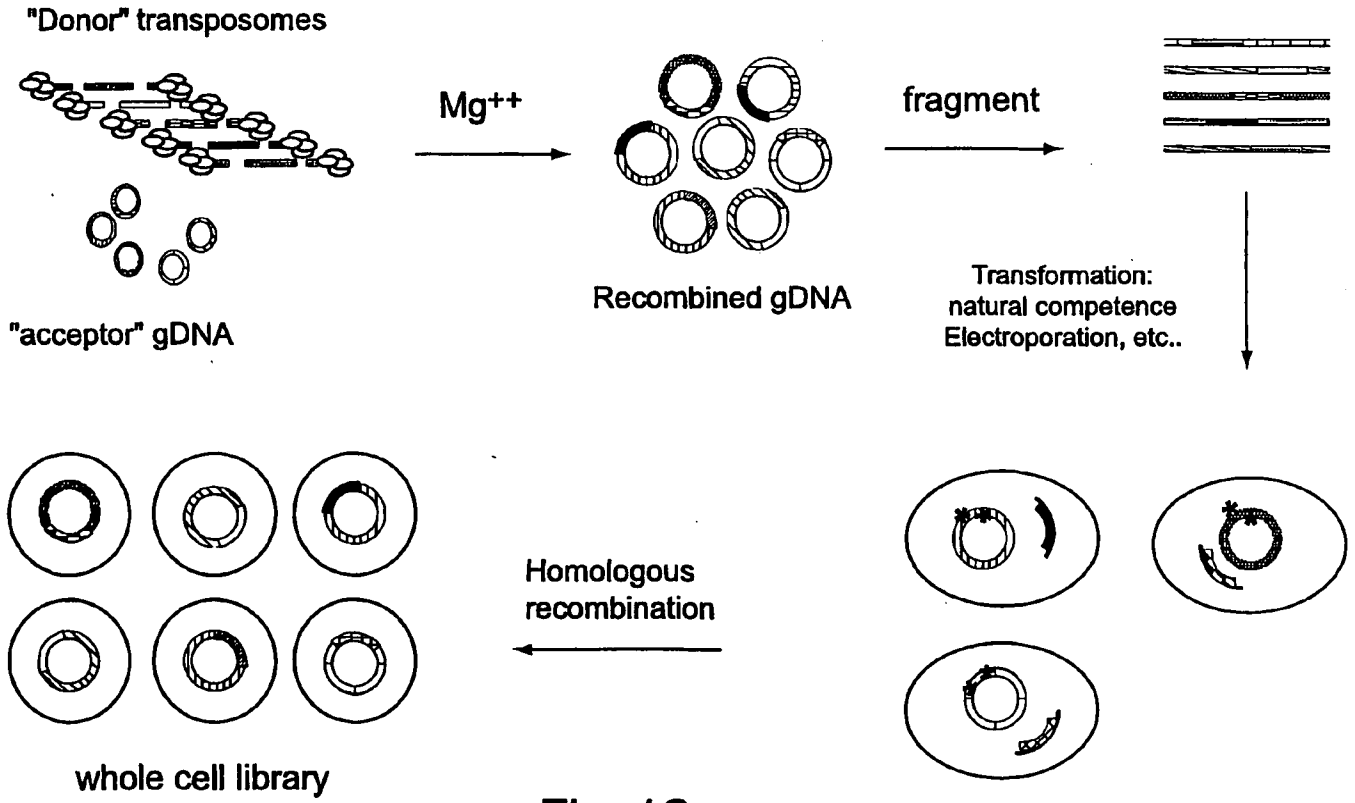


Fig. 4C

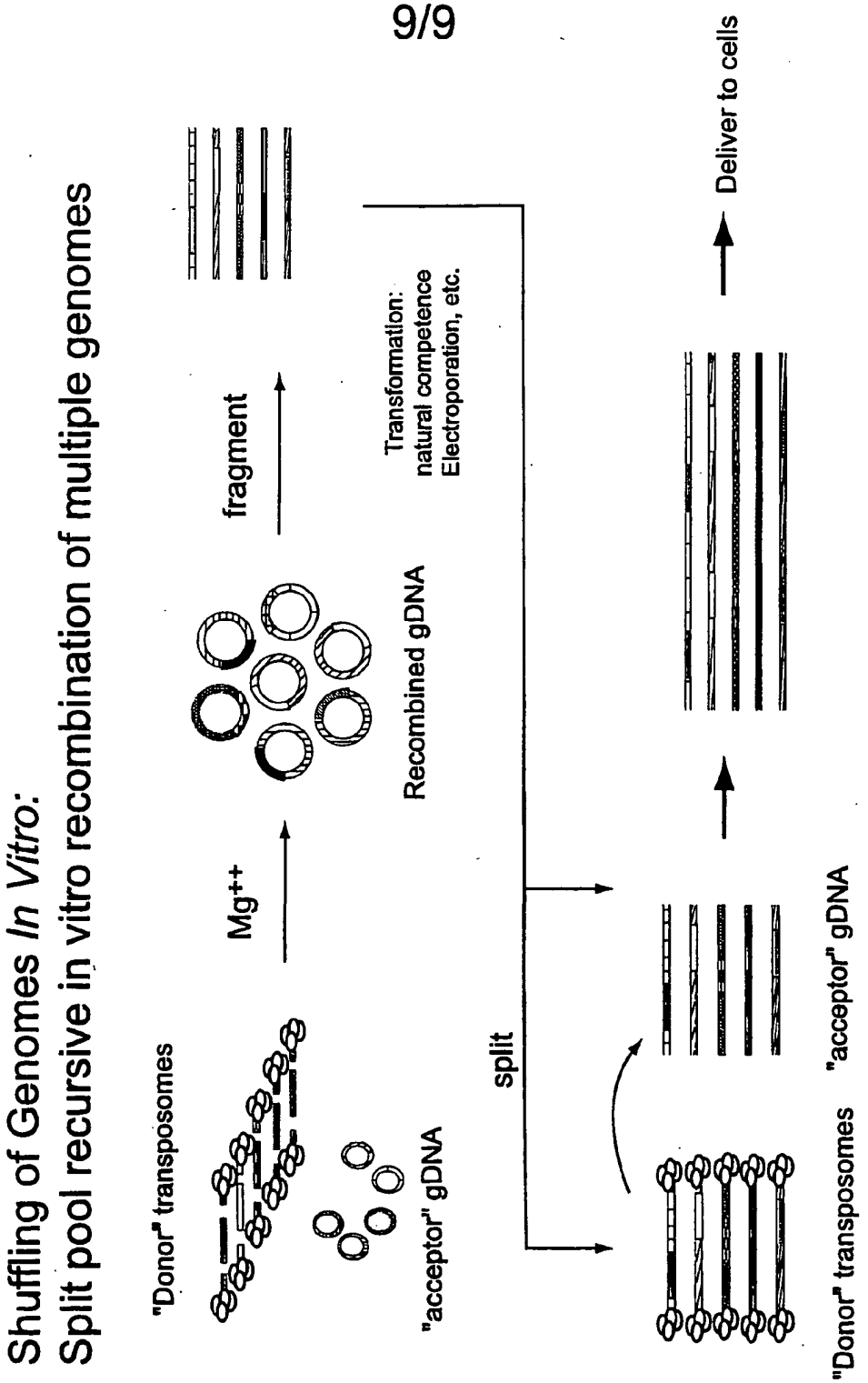

Shuffling of Genomes *In Vitro*:
Split pool recursive in vitro recombination of multiple genomes
"Donor" transposomes
"acceptor" gDNA
Mg++
Recombined gDNA
fragment
Transformation:
natural competence
Electroporation, etc.
split
Deliver to cells
"Donor" transposomes
"acceptor" gDNA

Fig. 4D

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
24 January 2002 (24.01.2002)

PCT

(10) International Publication Number
WO 02/06469 A2

(51) International Patent Classification[7]: C12N 15/10

(21) International Application Number: PCT/US01/22640

(22) International Filing Date: 18 July 2001 (18.07.2001)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

60/218,921	18 July 2000 (18.07.2000)	US
60/219,085	18 July 2000 (18.07.2000)	US
09/692,732	19 October 2000 (19.10.2000)	US
09/691,873	19 October 2000 (19.10.2000)	US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier applications:

US	60/218,921 (CIP)
Filed on	18 July 2000 (18.07.2000)
US	60/219,085 (CIP)
Filed on	18 July 2000 (18.07.2000)
US	09/691,873 (CIP)
Filed on	19 October 2000 (19.10.2000)
US	09/692,732 (CIP)
Filed on	19 October 2000 (19.10.2000)

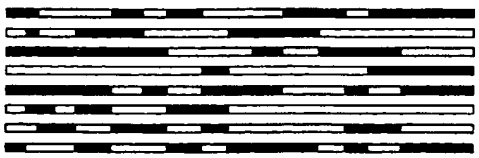
(71) Applicant *(for all designated States except US)*: **ENCHIRA BIOTECHNOLOGY CORPORATION** [US/US]; 4200 Research Forest Drive, The Woodlands, TX 77381 (US).

(72) Inventors; and

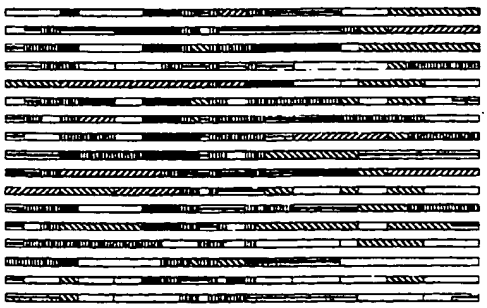
(75) Inventors/Applicants *(for US only)*: **COCO, Wayne, M.** [US/US]; 43 Cornerbrook Place, The Woodlands, TX 77381 (US). **ENCELL, Lance, P.** [US/US]; 58 Brooksedge Circle, The Woodlands, TX 77382 (US). **ARENSDORF, Joseph, J.** [US/US]; 27225 Orth Lane, Oak Ridge North, TX 77385 (US).

[Continued on next page]

(54) Title: METHODS OF LIGATION MEDIATED CHIMERAGENESIS UTILIZING POPULATIONS OF SCAFFOLD AND DONOR NUCLEIC ACIDS



A



B

(57) Abstract: The present invention is drawn to a method for forming at least one chimeric polynucleotide, methods for directed evolution, chimeric polynucleotides and libraries of chimeric polynucleotides. One method comprises contacting a first population of single-stranded oligonucleotides wherein the oligonucleotides share minimal complementarity with each other with a second population of oligonucleotides, under conditions wherein the oligonucleotides of the first and second populations hybridize to each other, forming at least one hybridized complex, comprising at least one polynucleotide from the first population hybridized to at least two oligonucleotides from the second population. Single-stranded regions are filled in using polymerase. The filled-in hybridized complex is treated such that the adjacent nucleic acids are ligated, forming at least one chimeric polynucleotide.

WO 02/06469 A2

(74) **Agents: ELMORE, Carolyn, S.** et al.; Hamilton, Brook, Smith & Reynolds, P.C., 530 Virginia Road,, P.O Box 9133, Concord MA 01742-9133 (US).

(81) **Designated States** *(national)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU, ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, TZ, UA, UG, US, UZ, VN, YU, ZA, ZW.

(84) **Designated States** *(regional)*: ARIPO patent (GH, GM, KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE, TR), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

METHODS OF LIGATION MEDIATED CHIMERAGENESIS UTILIZING POPULATIONS OF SCAFFOLD AND DONOR NUCLEIC ACIDS

RELATED APPLICATION

This application is a continuation-in-part of United States Application No. 09/692,732 filed October 19, 2000 and United States Application No. 09/691,873 filed October 19, 2000. This application also claims the benefit of United States Provisional Application No. 60/219,085, filed July 18, 2000 and United States Provisional Application No. 60/218,921 filed July 18, 2000. The teachings of all the above-referenced applications are hereby incorporated by reference in their entities.

BACKGROUND OF THE INVENTION

Genetic improvements occur more frequently when the generation of mutations is coupled with genetic recombination. Recombination between similar but non-identical polynucleotide targets allows for the consolidation of favorable mutations that appear on separate copies of the target, as well as the elimination of detrimental mutations (Harayama, S. *Trends Biotechnol.*, 16:76-82 (1998)). The effect of genetic recombination on the fixing of multiple beneficial mutations is noticeable when comparing sexually and asexually replicating organisms. Although single mutation rates are generally similar for sexually and asexually replicating organisms, juxtaposing these rare mutation events speeds up the evolutionary process dramatically. It is this ability to combine beneficial mutations and rapidly eliminate deleterious but not lethal mutations that enables sexually replicating organisms to evolve at a faster rate than asexually replication organisms. The reduction in evolutionary potential in asexually replicating populations is known as Müller's ratchet (Müller, H., *Mut. Res. 1*:2-9, 1964). The process of altering genetic functions through generation of mutants, and/or chimeric genetic recombinants,

coupled with selection and/or screening is termed "directed evolution."

The ability to generate a chimeric polynucleotide is fundamental to the process of directed evolution. Chimeric polynucleotides can result from recombination between two or more parent polynucleotides. To date, various strategies have been described to accomplish *in vitro* recombination. These include the following: "sexual PCR" (Stemmer, W. *Nature, 370*:389-391, 1994; United States Patent Nos. 5,605,793 and 5,811,238), which utilizes fragments cleaved from two or more parent double-stranded polynucleotides to form mutagenized double-stranded polynucleotides; "StEP" (Zhao, H. *et al., Nat. Biotechnol. 16*:258-61, 1998), which is characterized by multiple rounds of incomplete elongation of primers on variant templates; and the "RACHITT"™ method (Coco, W. *et al., Nat. Biotechnol. 19*:354-359, 2001), which typically uses the strategy of hybridizing fragments from one or more parents polynucleotides to a transient template, treating the overlaps and gaps enzymatically to yield a linear final product, and then destroying the original template prior to cloning.

These methods of directed evolution can form libraries of "chimeric" polynucleotides, so called because they include recombined sequences from more than one parent gene. The library of chimeric products, however, for sexual PCR and StEP, generally represent a limited and biased sampling of all potential chimeric products. These deficiencies are, in part, a result of limitations inherent in the essential PCR step in each of these methods. Moreover, these methods can suffer from "blind spots" in the gene or polypeptide of interest, where exchanges between parental DNA from two or more sources is rare or nonexistent due to the manner in which the DNA is fragmented or because regions of homology of a certain size are generally required to allow homologous recombination between the parental DNA. Although RACHITT™ overcomes these deficiencies to generate a broad library of chimeric polynucleotides, other distinct methods capable of producing libraries of chimeric polynucleotides of differing complexity would enhance the progress of directed evolution.

SUMMARY OF THE INVENTION

The methods of the present invention facilitate the generation of chimeric polynucleotides and do not require hybridizing donor fragments to a target- or full-length template. Rather, a first population of hybridizing donor fragments, *e.g.*, oligonucleotides, are assembled using a second population of scaffold fragments, *e.g.*, oligonucleotides, to form a double-stranded chimeric polynucleotide in which one or both strands are chimeric. One strand of the chimeric double-stranded polynucleotide can comprise, for example, scaffold fragments and regions between the scaffold fragments that were filled-in during the process; the opposite strand can also comprise donor fragments and regions between the donor fragments that were filled-in during the process. Because the chimeragenesis process of the present invention does not rely upon a contiguous, full-length template, it is unnecessary to modify a template to facilitate its removal.

In one embodiment, the invention is directed to a method for forming a chimeric polynucleotide including the steps of: contacting a population of single-stranded scaffold fragments with a population of donor fragments under conditions such that at least one scaffold fragment hybridizes to at least two donor fragments at distal regions of the scaffold fragment; treating the hybridized complexes such that single-stranded regions of the hybridized complex are filled-in; and treating the filled-in hybridized complexes such that adjacent fragments are ligated, forming a chimeric polynucleotide. In a particular embodiment, the method can also include the step of trimming flaps. Scaffold fragments can contain sequences of from about 10 to about 1000 nucleotides in length, preferably from about 25 to 100 nucleotides in length. Also, scaffold fragments can be derived from a single strand of a parent polynucleotide. Donor fragments can contain sequences of about 10 to about 1000 nucleotides in length, and they can be single-stranded. Donor fragments can be derived from a single strand of a parent polynucleotide. In a particular embodiment, scaffold and donor fragments hybridize to each other under conditions of low stringency. The population of scaffold fragments can be produced synthetically, or they can be produced by cleaving a polynucleotide of interest that is a full-length cDNA. The population of scaffold or donor fragments can include a fragment with

at least one region of random sequence. The method can further include a step of preparing at least one single-stranded population of scaffold fragments, derived from a randomly fragmented single-stranded polynucleotide of interest. In one embodiment, the populations of scaffold and donor fragments are sufficient to form a full-length chimeric polynucleotide. In a particular embodiment, the invention includes the step of screening or selecting at least one chimeric polynucleotide having desired characteristics. In another aspect, the invention is directed to chimeric polynucleotides prepared according to the methods described herein.

In another aspect, the invention is directed to a library of chimeric polynucleotides prepared according to the methods described herein. The library can be such that the majority of the chimeric polynucleotides contain at least 3 crossover sites. The library can contain at least one chimeric polynucleotide which contains the number of crossovers approaching the theoretical limit. The library can contain at least five chimeric polynucleotides which contains the number of crossovers approaching the theoretical limit.

In another embodiment, the invention is directed to a method for forming at least one double-stranded chimeric polynucleotide having desired characteristics including the steps of: contacting a population of scaffold fragments derived from a template polynucleotide with a population of donor fragments under conditions such that fragments of the scaffold and donor populations can hybridize to each other; forming at least one hybridized complex comprising at least one scaffold fragment hybridized to at least two donor fragments; treating the hybridized complex such that single-stranded regions of the hybridized complex are filled-in; treating the filled-in hybridized complex such that adjacent fragments are ligated, thereby forming a double-stranded chimeric polynucleotide. In one embodiment, the invention also includes the steps of trimming flaps and/or screening or selecting at least one double-stranded chimeric polynucleotide having desired characteristics. Scaffold fragments can contain sequences that are at least about 25 percent as long as a gene of interest. Scaffold and/or donor fragments can contain sequences of from about 25 to about 1000 nucleotides in length. In one embodiment, the donor fragments are single-stranded. Donor fragments can be such that they are derived from a single

strand of a parent polynucleotide. In one embodiment, the single-stranded regions are filled in using a polymerase. In one embodiment, the hybridized fragments are ligated using *Taq* DNA ligase or T4 DNA ligase. In a particular embodiment, the steps of hybridizing, filling in and ligating are repeated, such that one or more chimeric polynucleotides is used to generate the populations of scaffold or donor fragments. In one aspect, at least one of the fragments of the scaffold or donor populations contains at least one region of random sequence.

In another embodiment, the invention is directed to a method for preparing a population of scaffold fragments, including the steps of: amplifying an oligonucleotide of interest in a polymerase chain reaction, such that the 5' terminus of a first primer contains a 5' phosphate and the 5' terminus of a second primer is devoid of a 5' phosphate; contacting the amplified oligonucleotide with lambda exonuclease under conditions wherein oligonucleotides having a 5' phosphate are digested, leaving single-stranded oligonucleotides; and fragmenting the single-stranded oligonucleotides, thereby preparing a population of scaffold fragments.

In another embodiment, the invention is directed to a method for forming a chimeric polynucleotide including the steps of: treating a library of oligonucleotide fragments derived from a parent polynucleotide of interest and allelic variations thereof, wherein the population of fragments comprises a first population of oligonucleotides derived from one strand of the parent polynucleotide and allelic variations thereof and oligonucleotides of a second population wherein oligonucleotides are synthesized *in vitro* and derived from the other strand of the known parent polynucleotide and allelic variations thereof under conditions such that oligonucleotides of the first population can hybridize to oligonucleotides of the second population to form a gapped homoduplex; treating the gapped homoduplex with a polymerase, wherein polynucleotide strand extension produces a double-stranded polynucleotide comprising at least one nicked strand; and treating the nicked polynucleotide with a ligase, thus forming a full-length polynucleotide. In a particular embodiment, the invention is directed to a method of forming a single-stranded chimeric polynucleotide according, such that the oligonucleotides of the second population do not contain a 5' phosphate group, and includes the step of

removing the oligonucleotides of the second population after ligation. In a different embodiment, a single-stranded chimeric polynucleotide is formed using oligonucleotides of the second population that do not contain a 3' hydroxyl group. In the cases where a single-stranded chimeric polynucleotide is formed, scaffold fragments can be removed from the single-stranded chimeric polynucleotide after the ligation step. Single-stranded chimeric polynucleotides can be amplified in a nucleic acid amplification reaction to thereby produce more than one copy of a double-stranded chimeric polynucleotide. In one embodiment, at least one self-priming heteroduplex is a gapped heteroduplex including single-stranded sequences separated by double-stranded sequences. The gapped homoduplex can be full length. In one embodiment, the known parent sequence is from about 1 kilobase to about 5 kilobases in length. In another embodiment, the known parent sequence is from about 2 kilobases to about 25 kilobases in length. One aspect of the invention includes an additional recombination step between the chimeric polynucleotide and a parent molecule or allelic variation thereof.

In one embodiment, the invention is directed to a library of chimeric polynucleotides comprising more than one chimeric polynucleotides formed according to the methods described herein. The oligonucleotides of the second population can be derived from regions of sequence identity between parent polynucleotides and allelic variations thereof. In one embodiment, the gapped homoduplex can contain polymorphic sites in at least one double-stranded region of the homoduplex. In another embodiment, the gapped homoduplex can contain at least one polymorphic site in the gapped region of the gapped homoduplex.

In another embodiment, the invention is directed a method for directed evolution including the steps of: forming a library of chimeric polynucleotides by: contacting a first population of oligonucleotides with a second population of oligonucleotides, wherein the sequences of the first and second oligonucleotide populations are complementary to one another, under conditions such that oligonucleotides of the first population can hybridize to oligonucleotides of the second population to form a gapped homoduplex; treating the gapped homoduplex with a polymerase, such that polynucleotide strand extension produces a nicked

polynucleotide; treating the nicked polynucleotide with a ligase, such that nicks are ligated; and screening the library of chimeric polynucleotides for a characteristic of interest. In one embodiment, the oligonucleotides of the first population and the oligonucleotides of the second population are derived from a known polynucleotide of interest. In one aspect, the steps are repeated using the chimeric polynucleotide as the known polynucleotide of interest in the subsequent round of directed evolution. In a particular embodiment, the steps are repeated from about 2 to 50 times using a screened population of chimeric polynucleotides as the parent polynucleotides used to generate scaffold and donor fragments in a subsequent round of directed evolution. In one embodiment, the oligonucleotides of the second population do not contain 5' phosphate groups. In another embodiment, the oligonucleotides of the second population do not contain 3' hydroxyl groups. In a particular embodiment, the screening step includes screening the function of the transcribed and/or translated products of the library of chimeric polynucleotides. One aspect of the invention involves cloning the library of chimeric polynucleotides into a suitable vector prior to the screening step.

In a particular embodiment, the methods for directed evolution described herein include: cloning the chimeric polynucleotides into expression vectors; transforming a suitable cell line with the cloned chimeric polynucleotides; inducing expression of the cloned chimeric polynucleotide; assaying the expressed product for a characteristic of interest; and selecting the chimeric polynucleotide that expressed products with an improved characteristic of interest. In another embodiment, the methods for directed evolution described herein include: transcribing and translating the chimeric polynucleotide *in vitro*; assaying the transcribed and translated products for a characteristic of interest; and selecting the chimeric polynucleotide that lead to transcribed and translated products with an improved characteristic of interest.

BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing and other objects, features and advantages of the invention will be apparent from the following more particular description of an embodiment of the invention, as illustrated in the accompanying drawings.

Figure 1 is a schematic diagram of one embodiment of the present invention.

Figures 2A and 2B depict synthetic genes and oligonucleotides. 2A. Alignment of TCI and PCI sequences altered to represent common *E. coli* codon usage. SNPs, DiPs and TriPs are as indicated. 2B. Oligonucleotides used for *in vitro* recombination of TCI and PCI genes. Restriction sites for cloning are underlined. Degenerate positions are shown in alternative bases in parentheses. Permutations of DiPs and TriPs are underlined. Note that the primers anneal to the genes in panel A and represent only the top strand.

Figures 3A and 3B are alignments of the human and mouse EGF coding sequences. 3A. Alignment of the unmodified human and mouse EGF coding sequences (with flanking engineered sequences and restriction cleavage sites). Identical positions are marked by dashes in the mouse sequence. Amino acid residue polymorphisms are indicated below the alignment. 3B. Alignment of the genes after design modifications to minimize genetic differences without altering the information content of the encoded polypeptides.

Figures 4A-C are schematic diagrams depicting different shuffling strategies. 4A. PARSed DNA shuffling of the mouse and human EGF polymorphisms. 4B. PARSed DNA shuffling of EGF polymorphisms from five species (human, mouse, rat, horse and pig). 4C. Heteroduplex DNA shuffling of degenerate oligonucleotides.

Figures 5A and 5B are schematic diagrams representing PARSed DNA shuffling products. 5A. Two-gene PARSed shuffling. Black blocks contain only human nucleotide polymorphisms. White blocks contain only mouse polymorphisms. 5B. Five-gene PARSed shuffling. Blocks containing codons from each of the five mammalian species are uniquely shaded. Regions containing polymorphisms that cannot be assigned to a single parent are left unshaded. Unambiguous crossovers in such regions are indicated by vertical lines.

Figure 6 is a graph showing the frequency of reassortment of polymorphisms. DNA sequence information for 8 unselected clones indicated representation of each allele at every polymorphic position. The number of human vs. mouse alleles at any one position ranged from 2 to 6 (fraction = 0.25 to 0.75) and clustered near the theoretically ideal value of 0.5.

DETAILED DESCRIPTION OF THE INVENTION

The methods of the present invention facilitate the generation of chimeric polynucleotides and do not require hybridizing donor fragments to a target- or full-length template. Rather, a first population of hybridizing donor fragments, *e.g.*, oligonucleotides, are assembled using a second population of scaffold fragments, *e.g.*, oligonucleotides, to form a double-stranded chimeric polynucleotide in which one or both strands are chimeric. One strand of the chimeric double-stranded polynucleotide can comprise, for example, scaffold fragments and regions between the scaffold fragments that were filled-in during the process; the opposite strand can also comprise donor fragments and regions between the donor fragments that were filled-in during the process. Because the chimeragenesis process of the present invention does not rely upon a contiguous, full-length template, it is unnecessary to modify a template to facilitate its removal.

"Chimeric polynucleotides", as used herein, contain nucleotide sequences from multiple related sequences or otherwise similar polynucleotides, referred to herein as "parent polynucleotides." "Full-length," as used herein to describe polynucleotides, is a relative term meaning the product is about the same length as the parent polynucleotide. In one embodiment, the scaffold is made up or otherwise designed, generated or derived of fragments from one strand, *e.g.*, the top strand, of a parent polynucleotide, *e.g.*, a template polynucleotide. In another embodiment, the scaffold is formed without reference to a particular strand of a parent polynucleotide, but the scaffold fragments are nonetheless complementary to the donor fragments.

The methods described herein comprise process steps involved in the formation of chimeric polynucleotides. Reference is now made to Figure 1 which depicts schematically the steps utilized by one embodiment of the present invention

in forming the double-stranded chimeric polynucleotide, wherein a population of scaffold fragments are used to assemble the hybridizing population of donor fragments. A polynucleotide of interest, *e.g.*, a gene, 10 is used to prepare a population of single-stranded scaffold fragments 20. A population of donor fragments 30 is assembled into hybridization complexes 40 with the scaffold fragments. In some cases, overlaps occur between donor fragments and/or scaffold fragments, thus creating "flaps" 50. The term "flaps" is intended to include the unhybridized terminal portions of a fragment that is otherwise hybridized to another fragment. Overlaps can occur between hybridized scaffold fragments or hybridized donor fragments. In other cases, regions between the hybridized fragments remain single-stranded, thus creating "gaps" between the fragments 60. Flaps can be trimmed and gaps can be filled-in prior to the generation of a contiguous chimeric polynucleotide 70. A contiguous double-stranded chimeric polynucleotide can be generated by ligating the assembled oligonucleotides 80. The method of the present invention can further include repeating the method using at least one chimeric polynucleotide or fragment thereof as the scaffold fragments or donor fragments.

Figure 4 is a schematic diagram depicting different shuffling strategies. Figure 4A depicts PARtially Scaffolded (PARSed) DNA shuffling of the mouse and human EGF polymorphisms. Three degenerate 5'-phosphorylated top strand (TS1-3) and two non-phosphorylated partial scaffold (PS1/2) oligonucleotides were synthesized to contain all the amino acid polymorphisms of the parental mouse and human *EGF* genes and silent modifications. Arrows indicate the position and number of alternative codons. Arrows opposite gaps in the top strand indicate SNPs that were incorporated by primer extension, using the scaffold oligonucleotides as templates. Boxes indicate where alternative codons were synthesized in separate reactions in order to minimize degeneracy. Dotted vertical lines indicate homoduplex base pairing between top strand degenerate positions and their complementary bases in the scaffolds. Bold numbers above or below each oligonucleotide indicate its length in nucleotides. Numbers between oligonucleotides indicate the number of nucleotides available for hybridization up to the first degenerate position. Spaces between top strand oligonucleotides represent

nicks, except where underlined numbers indicate gap lengths. Figure 4B shows PARSed DNA shuffling of EGF polymorphisms from five species (human, mouse, rat, horse and pig). The two gray arrows indicate inclusion of amino acids not present in the parental genes. Figure 4C shows a method for heteroduplex DNA shuffling of degenerate oligonucleotides. Oligonucleotides with polymorphisms from two parental genes were annealed to a full length template representing either of the two parental genes. Arrows topped with an "x" indicate the heteroduplex mismatches closest to the oligonucleotide ends. As used herein, "homoduplex" polynucleotides refer to hybridized strands that contain only Watson-Crick base pairs, *i.e.*, they do not contain "mismatches." "Heteroduplexes" are hybridized polynucleotide molecules that contain at least one mismatch.

The chimeric polynucleotides described schematically in Figures 1 and 4 include crossovers. As used herein, "crossover" refers to an event that leads to strand switching. As used herein, "strand switching" describes a nucleotide sequence such that the sequence is identical to a reference polynucleotide up to the "switch" or "crossover" site, and the sequence downstream of the crossover site is identical to a different reference polynucleotide. Strand switching is a description of nucleotide sequence and is not necessarily indicative of a physical switching of strands; see, for example, Figures 3 and 5 which depict chimeric sequences containing crossovers.

In embodiments, the population of scaffold fragments is derived from one or more allelic versions of the gene of interest. As used herein, "allelic version" refers to a polynucleotide with a sequence similar to a reference polynucleotide, *e.g.*, a wild-type gene. The allelic version typically has a different sequence at one or more "polymorphic sites" with respect to the reference polynucleotide. As used herein, "polymorphic sites" refer to those positions in a sequence where, in a population of related polynucleotides, more than one sequence occurs. As used herein, "degeneracy" refers to either the number of allelic variants at a polymorphic site, or the number of polymorphic sites contained in a nucleotide sequence; a higher level of degeneracy corresponds to a greater number of variants possible at a polymorphic site or a greater number of polymorphic sites in a particular sequence. In contrast,

among the population of polynucleotides, the sequences are identical at non-polymorphic sites. Polymorphic sites can exhibit single nucleotide polymorphisms (SNPs), dinucleotide polymorphisms (DiPs) and trinucleotide polymorphisms (TriPs) or combinations thereof. Expected frequencies of crossovers can be calculated based on the fold degeneracy at a particular polymorphic site. For example, if the theoretical limit of crossover events is reached, one would expect to have an equal chance (determined by the input of the donor and scaffold fragments) of having a particular allelic variant at a particular polymorphic site irrespective of the allelic variant present at a different polymorphic site. For example, if a two-fold degenerate polymorphic site is included in a shuffling reaction such that an equal number of each variant is used to generate scaffold and donor fragments, it would be expected that 50% of the resultant chimeric polynucleotides would contain each allelic variant. This 50% value is independent, *i.e.*, does not display "genetic linkage," of the specific allelic variant at a different polymorphic site.

A population of scaffold fragments can be contacted with a population of donor fragments. The interactions between donor fragments and scaffold fragments occur by the process of hybridization. Thus, some degree of complementarity between scaffold fragments and donor fragments must exist to allow for such interactions. Further, since the interactions between scaffold and donor fragments are dependent on base-pairing, in a particular embodiment, scaffold fragments are derived from a polynucleotide strand complementary to the strand from which the donor fragments are derived. In one embodiment, scaffold fragments are derived from a reference polynucleotide or "template" polynucleotide. In another embodiment, scaffold fragments are derived from the strand of the allelic version that is complementary to the strand that is used to generate the donor fragments.

The scaffold fragments typically comprise single-stranded molecules, having minimal complementarity with each other. As used herein, "minimal complementarity" between scaffold fragments means that a scaffold fragment will hybridize with other fragments, *e.g.*, donor fragments, to form a duplex with a higher melting temperature than a duplex formed by the scaffold fragment hybridizing to another scaffold fragment. In a preferred embodiment, the single-stranded scaffold

fragments are prepared, synthesized, designed, generated or otherwise derived from one strand, *e.g.*, the top strand of a polynucleotide of interest. Single-stranded scaffold fragments can be prepared by nicking a single-stranded polynucleotide of interest or by denaturing a nicked double-stranded polynucleotide. Single-stranded polynucleotides can be made by denaturing double-stranded polynucleotides. "Denaturing," as used herein, refers to the process of physically separating the single strands of nucleic acid by disrupting base-pairing interactions between complementary strands. Such methods of denaturing double-stranded polynucleotides are well known in the art.

Scaffold fragments can be made, for example, using enzymatic techniques, physical techniques or chemical synthesis techniques. For example, single-stranded polynucleotides can be synthesized by PCR amplification of a polynucleotide of interest using primers wherein one primer contains a terminal 5' phosphate and the other primer does not contain a terminal 5' phosphate. The amplified product can then be treated such that nucleic acid having a terminal 5' phosphate or molecules devoid of a terminal 5' phosphate are preferentially destroyed. In one embodiment, the amplified product is treated with lambda exonuclease to degrade the phosphorylated strand. Single-stranded polynucleotides can also be made by inserting a polynucleotide of interest into M13 phage and performing first strand synthesis using methods well known in the art. The single-stranded polynucleotides can be fragmented by enzymatic, chemical, or physical techniques known in the art in order to generate single-stranded scaffold fragments.

The scaffold fragments can be generated from a larger polynucleotide and treated to form fragments or single-stranded fragments. In embodiments, double-stranded polynucleotides are fragmented such that fragments of one strand form the population of donor fragments and fragments of the complementary strand form the population of scaffold fragments. In this embodiment, the double-stranded polynucleotides that are fragmented are, preferably, different allelic versions of each other.

The scaffold fragments can be of any length which is less than that of a full-length polynucleotide of interest, *e.g.*, less than the full-length of the corresponding

wild-type gene. Preferably, scaffold fragment are considerably shorter than the full-length polynucleotide, most preferably not more than 30% of its length. In one embodiment, the scaffold fragments can be from about 20 to about 1500 nucleotides in length. In another embodiment, the scaffold fragments can be from about 25 to about 1000 nucleotides in length or from about 100 to about 1000 nucleotides in length. The scaffold fragments can be at least about 40 nucleotides in length, at least about 100 nucleotides in length, or at least about 1000 nucleotides in length. The scaffold fragments can be less than about 25 percent of the desired length of the chimeric polynucleotide products, or about 15 or 20 percent or less of the desired length of the chimeric polynucleotide products. Without wishing to be bound by theory, while the use of longer scaffold fragments can facilitate the formation of target length chimera in the absence of thermocycling or multiple rounds of annealing and denaturing, shorter scaffold fragments can facilitate the number of crossovers. As used herein, "target-length chimera" refers to the approximate length of a hypothetical chimeric polynucleotide having the desired properties. Target length can be estimated based on the length of the polynucleotide of interest or reference polynucleotide as described herein.

The scaffold fragments allow for the assembly of donor fragments into an ordered duplex with at least one scaffold fragment. Typically, the scaffold fragments are selected such that they are related to the parent polynucleotides of interest, *e.g.*, genes, which are allelic versions of each other, that are used to generate the donor fragments. In another embodiment, the scaffold fragments are derived from a reference polynucleotide of interest (a "template") and the donor fragments are derived from allelic versions of the template or a combination of allelic versions and the template. In a particular embodiment, the scaffold fragments are derived from a particular strand of a duplex polynucleotide. For example, the scaffold fragments can be derived from the sense or top strand, and the donor fragments can be derived from the antisense or bottom strand. The polynucleotide of interest can comprise a gene, either a genomic copy or cDNA (or intronless) copy. The polynucleotide of interest can comprise more than one coding sequence. For

example, the polynucleotide of interest can comprise an operon including regulatory regions, either as a single contiguous molecule or as more than one molecule.

The nucleic acids for use as either scaffold or donor fragments can be synthetically manufactured or isolated from any suitable source of nucleic acid. The scaffold or donor fragments of the present invention can comprise DeoxyriboNucleic Acid (hereinafter "DNA"), or RiboNucleic Acid (hereinafter "RNA") DNA or RNA can comprise natural bases, *e.g.*, adenine, thymine, cytosine, guanine or uracil; analog bases, *e.g.*, inosine, bromouracil or nitroindole; chemically altered bases, *e.g.*, biotin labeled or digoxygenin labeled bases; or a combination thereof provided that the resulting double-stranded chimeric polynucleotide can be replicated. Scaffold fragments can be such as the cannot be ligated, *e.g.*, they lack 5' phosphate groups, or they can be such that they can not be extended, *e.g.*, they lack 3' hydroxyl groups.

Further, polynucleotides used to generate the scaffold or donor fragments, or the fragments themselves can be isolated from an organism, such as, for example, a eubacterial, archeal, eukaryotic or viral organism. These organisms can be amplified, enriched or isolated and grown in culture, or can be used directly from environmental sources. Environmental sources include soil samples, water samples from fresh water sources or salt water sources, polluted sites, waste treatment sites and sources including extreme condition sources such as permafrost sources, high altitude sources, high pressure sources and geothermal sources such as volcanic sources, hot springs and hydrothermal vent sources. Sources of nucleic acid also include tissue or bodily fluid samples from an organism, such as a human samples and include human genomic DNA. The nucleic acid of a tissue or bodily fluid sample can include nucleic acid of the organism, such as chromosomal, episomal or transcribed nucleic acid, or can be nucleic acid of the flora, such as fungal, bacterial, viral or parasitic organisms present in the sample. The sample can further be fresh, fossil or archival.

It is understood that the nucleic acid isolated from these sources can be produced in the form of a genomic or cDNA library using methods well known in the art. In the case of cDNA, RNA or preferably polyA$^+$ RNA or mRNA is isolated from a sample, and converted into double-stranded DNA (cDNA) according to

standard methods, well known in the art. In one embodiment, a cDNA library is prepared from a sample of interest that expresses the desired phenotype. In another embodiment, the cDNA library can be enriched for sequences of interest prior to use as oligonucleotides. The cDNA library can be subjected to subtractive hybridization against a suitable sample of nucleic acid using subtractive hybridization techniques well known in the art. A suitable sample of nucleic acid includes, for example, nucleic acid from a reference strain of bacteria. In one embodiment, sequences that are common between the cDNA library and the sample nucleic acid are allowed to hybridize to each other and double-stranded nucleic acids are then removed from the pool. In this way, sequence present in multiple copies and sequences that are common between the two populations are removed, effectively enriching for low abundance or unique sequences. For example, a library of donor fragments prepared according to the method as described in "Generating Single-Stranded Oligonucleotide Libraries with Minimal Complementarity and Uses Therefore" by Joseph J. Arensdorf and Wayne M. Coco, United States Application No. 09/691,873 filed October 19, 2000.

The scaffold or donor fragments of the present invention can be isolated from any suitable source of oligonucleotides as described herein. Methods of choosing and/or isolating nucleic acids from suitable sources of nucleic acid are well known in the art. In another embodiment of the present invention, the scaffold or donor fragments (or both) can be produced *in vitro* using enzymatic or chemical means. Methods of *in vitro* production of nucleic acid sequence are well known in the art.

The scaffold or donor fragments can include one or more regions with functional characteristics or structural motifs of the parent polynucleotides. The scaffold or donor fragments can comprise all or a portion of a region with functional characteristics or structural motifs. These regions can include nucleic acid structural motifs, protein binding domains, metal binding domains, nucleic acid binding domains, domains with enzymatic activity, or fragments of these domains. These regions can include ribozymes, deoxyribozymes, promoters, enhancers, origins of replication, open reading frames, or fragments thereof. These regions can encode

aptamers, wherein aptamers are small single- or double-stranded DNA or RNA molecules that bind specific molecular targets (Bock *et al.*, *Nature 355*:564-566, 1992; Ellington and Szostak, *Nature 346*:818-822, 1990; Werstuck and Green, *Science 282*:296-298, 1998).

The scaffold or donor fragments of the present invention can also include regions of sequence that are not known to have any particular function. These regions can be selected from any known source of nucleic acid sequence, including sequences synthesized *in vitro*, or these regions can be of random or partially random sequence. Partially random sequences can be generated by synthesizing a oligonucleotide based on a known sequence, except that a portion of the sequence is randomized (*e.g.*, randomizing the last 50 nucleotides), or wherein certain positions within the sequence are randomized (*e.g.*, randomizing particular codon(s) of a coding sequence) or wherein certain bases are randomized (*e.g.*, randomizing all adenines). These regions can further encode proteins or domains of proteins including folding structures or structural motifs; binding domains such as protein binding domains, metal binding domains, co-factor binding domains, lipid binding domains and nucleic acid binding domains; domains with enzymatic function; sites for allosteric or competitive inhibition and the like; or fragments of these domains. These regions can also include amino acid sequences that are not known to have any particular function or can be randomized amino acid sequence.

The parent polynucleotides can be fragmented while in double-stranded or single-stranded form. Preferred methods for cleaving, *e.g.*, fragmenting parent polynucleotides in order to generate populations of donor and scaffold fragments are those methods which produce fragments without particular sequence patterns. In one embodiment of the present invention, a population of fragments is created by randomly fragmenting parent polynucleotides.

The parent polynucleotides, scaffold fragments or donor fragments are generated using chemical, physical or enzymatic techniques. Chemical techniques of fragmenting polynucleotides can include techniques that utilize pH extremes, hydroxyl radical formation, chemical radical formation, chemical catalysis or a combination thereof. Methods of fragmenting polynucleotides by chemical

techniques can be used to generate defined or undefined ends. Techniques are well known in the art such that polynucleotides can be hydrolyzed after defined bases (*e.g.*, only after guanines), or hydrolyzed to generate undefined termini. For example, exposure of polynucleotides to extreme pH (*e.g.*, acidic pH or basic pH) can generate fragments with undefined termini. Additionally, hydroxyl radicals (*e.g.*, generated using Fenton or Udenfriend reagent) react with the deoxyribose in DNA, resulting in cleavage of the DNA strand. The result is near uniform cleavage at any base within a target polynucleotide, and the frequency of cleavage can be regulated. In addition to fragmenting polynucleotides by chemical techniques, physical techniques, such as heating, freezing, using ionizing radiation and shearing can be employed.

Yet another approach to creating a population of fragments involves the use of enzymatic techniques. These methods can include the use of any suitable enzyme such as a nucleic acid polymerizing enzyme or a nuclease. For example, a polymerase can be used to synthesize oligonucleotides of variable length. Where fragments are generated by parent polynucleotide-dependent synthesis, conditions of synthesis can be chosen such that the polymerase arbitrarily falls off the polynucleotide or otherwise terminates synthesis at arbitrary points along the polynucleotide. This approach allows for oligonucleotides to be generated with arbitrary sequence alterations (*e.g.*, "error-prone" methods). Another method for using polymerases to generate a fragmented population of oligonucleotides uses polymerases that are known to have exonuclease activity under conditions permitting exonuclease activity. Such enzymes include, for example, T4 DNA polymerase, PolI, PolIII, *Pfu* polymerase and Klenow polymerase.

Still another method for enzymatically generating a population of oligonucleotides with undefined termini involves removing bases or generating adducts in an oligonucleotide using techniques well known in the art. For example, specific bases in oligonucleotides can be removed or adducted by many well known chemical methods to result in either abasic sites or chemically altered bases. These sites can be produced, for example, between 15 and 5000 bases apart (Kunkel *et al.*, *Meth. Enzymol. 154*:367-382, 1987). Strand cleavage of the phosphodiester bond at

those modified sites can then be effected using chemicals such as piperidine, or enzymes such as abasic lyases or abasic endonucleases.

Another enzymatic method for creating a fragmented population of oligonucleotides uses endonucleases having sequence-specific recognition sites. Such enzymes are known as "restriction endonucleases" and are commercially available. A fragmented population of oligonucleotides can be generated by performing a limited or incomplete digestion of the parent polynucleotides. Additionally, oligonucleotides having undefined termini can be generated by using non-specific endonucleases such as mung bean exonuclease, S1 nuclease or DNase I. In another embodiment of generating oligonucleotides having undefined ends, exonucleases such as ExoIII or ExoVII can be used to non-specifically trim oligonucleotide sequences.

The fragmented population of oligonucleotides can include oligonucleotides containing random or partially random sequence. The population of fragments can include molecules generated using any one of the above described methods or combinations thereof. The term "random" as used herein is intended to reflect an absence of preselection. Such absence can be of any degree; it need not be a total absence of preselection, nor does the term indicate a requirement for an absence of preference or bias. The term can be used to describe populations of oligonucleotides, sequences, events, processes, states or conditions, or other such terms. Such compositions can range over a span of values and any one component can occupy any of these values. For example, a population of oligonucleotides that is generated by the digestion of two polynucleotides with a restriction enzyme is a "random population" when the particular oligonucleotides formed by the process are not preselected, for example, during a partial digestion. This is true even when the gene sequences are known and the restriction enzyme preferentially cleaves a particular site. Sequences can be random if at least one position in the sequence is not specifically defined (for example, if at least one position of an oligonucleotide could be and is either one of two or more nucleotides, *e.g.*, a polymorphic site). The randomly fragmented population of oligonucleotides can include oligonucleotides

wherein a portion of the oligonucleotides comprise random or partially random sequence as described herein.

In a preferred embodiment, the scaffold population is not randomly produced but is designed to optimize crossover events. Such a scaffold population can provide either chimeragenesis or a lack of chimeragenesis, but will correspond to the whole or a substantial, although not necessarily contiguous, length of the polynucleotide of interest, *e.g.*, gene of interest. For example, scaffold fragments can be synthetically produced to each contain complementary sequences, *i.e.*, termini, to two donor fragments. Thus, each terminus of the scaffold fragment can hybridize to a different donor fragment and each donor fragment terminus (with the exception of flaps) can hybridize to a different scaffold fragment terminus. The donor fragments can be randomly generated or specifically designed to introduce chimeragenesis. The scaffold can be designed to have identity to the termini of the donor fragments to provide the desired cross-overs and gaps to correspond to the desired mutations or chimeragenesis.

The population of scaffold fragments of the present invention includes oligonucleotides that are typically shorter than target length chimera. The target length *e.g.*, length of resulting double-stranded chimera, can be from about 50 to about 100,000 nucleotides in length. In particular embodiments, the target length can be from about 100 to about 50,000 nucleotides in length; from about 200 to about 10,000 nucleotides in length; from about 500 to about 5,000 nucleotides in length or from about 1,000 to about 3,000 nucleotides in length.

The population of donor fragments includes oligonucleotides from about 5 to about 50,000 nucleotides length. In more particular embodiments, the population of donor fragments includes oligonucleotides from about 10 to about 10,000 nucleotides in length, from about 15 to about 5,000 nucleotides in length, from about 20 to about 2,500 nucleotides in length, from about 25 to about 1,000 nucleotides in length, or from about 40 to about 200 nucleotides in length. The donor fragments can be at least about 40 nucleotides in length, at least about 100 nucleotides in length or at least about 1000 nucleotides in length.

The scaffold fragments guide the hybridizing donor fragments and form a double-stranded chimeric polynucleotide. Where the donor fragments are double-stranded molecules, they can be denatured prior to hybridization with the scaffold fragments. Methods of denaturing and annealing donor fragments sequences, are well known in the art. In a particular embodiment, the donor fragments are single-stranded and from the opposite strand compared to the scaffold fragments. "Opposite strand" refers to, for example, the donor fragments being derived from the antisense or bottom strand of a duplex polynucleotide when the scaffold fragments are derived from the top or sense strand. In a particularly preferred embodiment, the scaffold fragments are derived from the top strand, and the donor fragments are derived from the bottom strand. In another embodiment, the population of donor fragments or the scaffold fragments, or both share minimal complementarity with members of the same population. Minimal complementarity can be achieved by treating the members of the population such that they do not hybridize to each other. Alternatively, minimal complementarity can be achieved by selecting the members of the population such that they do not hybridize to each other, *e.g.*, forming the population by cleaving one strand of a polynucleotide of interest. It is clear that one of skill in the art can skew the availability of a given donor fragments to hybridize a scaffold fragment by including an oligonucleotide capable of hybridizing to said donor fragment. In one embodiment, a region of the scaffold is hybridized to complimentary sequences by providing oligonucleotides complimentary to a specific scaffold sequence. In this manner, a region of the scaffold can be specifically retained in the resultant double-stranded chimeric molecule. Conversely, defined oligonucleotides can be added in greater quantities to the population of donor fragments in order to preferentially hybridize the defined oligonucleotides to the scaffold at particular regions or positions in order to introduce desired mutations or in order to protect sequences on the scaffold from changes that might be introduced by the arbitrarily fragmented population of donor fragments.

The donor fragments can also be single-stranded, and in a particular embodiment, are derived from the opposite, *e.g.*, complementary, strand to the strand of a duplex polynucleotide from which the scaffold fragments are derived. As

used herein, "derived" refers to a sequence identical to a sequence contained in a reference polynucleotide except at polymorphic sites. Thus, when the donor and scaffold fragments contact each other, a hybridized complex can form, which generally comprises at least one donor fragment hybridized to at least one scaffold fragment. In a particular embodiment, the hybridized complex comprises at least two donor fragments hybridized to at least one scaffold fragment. Single-stranded regions remaining between adjacently hybridized fragments, herein referred to as "gaps," can be filled in, *e.g.*, using a polymerase. Where there is a 3' overhang, the overhang can be filled in by adding a primer that hybridizes at or near the free terminus of the 3' overhang such that polymerization during gap filling can proceed on the 3' overhang. Adjacently hybridized fragments can then be ligated to form a double-stranded polynucleotide comprising chimeric polynucleotides.

The present invention allows donor fragments of interest and scaffold fragments to be incorporated into a larger molecule to form one or more double-stranded chimeric polynucleotides. In one embodiment, polynucleotides that are not otherwise easily manipulated (*e.g.*, large polynucleotide chains), can be separately manipulated as oligonucleotides and rejoined by contacting the oligonucleotides with single stranded scaffold fragments to form a hybridized complex. For example, random mutagenesis using PCR is most effective on smaller DNA fragments, such as 1 kilobase or less in length. A large polynucleotide can be cleaved into fragments of about one kilobase, randomly mutagenized using PCR, and then denatured. Denatured and mutagenized fragments can be contacted with scaffold fragments to form a hybridized complex, filled in and ligated as described herein. The template scaffold can be derived from the original oligonucleotide, or can be modified as described herein. For example, the scaffold fragments can be mutagenized or can have added or deleted regions or domains as compared to the starting polynucleotide.

It is clear to one of skill in the art that the method of the present invention can be carried out under a range of reaction conditions and hybridization conditions. Conditions can be selected based on the amount of similarity or differences between the oligonucleotides and the template. In one embodiment of the present invention,

the donor fragments are hybridized or annealed to the scaffold fragments under conditions of low stringency.

A general description of stringency for hybridization and wash conditions is provided by Ausubel, F.M. *et al.*, *Current Protocols in Molecular Biology*, Greene Publishing Assoc. and Wiley-Interscience 1987, & Supp. 49, 2000, the teachings of which are incorporated herein by reference. Factors such as probe length, base composition, percent mismatch between the hybridizing sequences, temperature and ionic strength of reactions, hybridizations and washes influence the stability of nucleic acid hybrids. Thus, stringency conditions sufficient to allow hybridization of donor and scaffold fragments to form hybridization complexes can vary significantly and still allow for the generation of at least one chimeric polynucleotide. The energetics favoring hybridization indicate that longer stretches of homology are more favorable. Thus, when either short sequences are involved or there is limited potential for standard Watson-Crick base-pairing, hybridization conditions can be adjusted to a lower stringency to allow for hybridization. Typically, adjusting hybridization and wash conditions is done by, for example, adjusting the ionic strength of the reaction mixture or adjusting the temperature at which the hybridization is performed. In addition, certain purified proteins, such as the *E. coli* RecA protein, aid in homologous base pairing and can be included to facilitate hybridization of polynucleotide strands.

While not wishing to be bound by theory, typically, when two fragments anneal to form a hybridization complex, one or two single-stranded termini remain. These single-stranded termini can anneal to additional fragments from the mixture by altering hybridization conditions to favor the annealing of multiple fragments in a hybridization complex. To facilitate the hybridization of fragments having low homology, the donor and scaffold fragments can be allowed to anneal (hybridize) at 50°C. In another embodiment, the donor and scaffold fragments can be allowed to anneal at 60°C or at 70°C. To facilitate the hybridization of multiple donor fragments and scaffold fragments in a hybridization complex, the donor and scaffold fragment mixture can be held at the annealing temperature for at least about 30 seconds. In another embodiment, the donor and scaffold fragment mixture can be

held at the annealing temperature for at least about 1 minute, 2 minutes, 5 minutes, 15 minutes, 30 minutes, 1 hour, 5 hours, 10 hours or 24 hours. Combinations of annealing temperature and incubation time at the annealing temperature can be used to facilitate the formation of hybridization complexes comprising multiple donor and scaffold fragments.

Alternatively, conditions for stringency are as described in WO 98/40404, the teachings of which are incorporated herein by reference. In particular, examples of "highly stringent," "stringent," "reduced," and "least stringent" conditions are provided in WO 98/40404 in the Table on page 36. Examples of stringency conditions are shown in the table below which is from WO 98/40404. Highly stringent conditions are those that are at least as stringent as, for example, conditions A-F; stringent conditions are at least as stringent as, for example, conditions G-L; and reduced stringency conditions are at least as stringent as, for example, conditions M-R.

Stringency Condition	Oligonucleotide Hybrid	Hybrid Length (bp)‡	Hybridization Temperature and Buffer†	Wash Temperature and Buffer†
A	DNA:DNA	≥ 50	65°C; 1xSSC -or- 42°C; 1xSSC, 50% formamide	65°C; 0.3xSSC
B	DNA:DNA	<50	T_B*; 1xSSC	T_B*; 1xSSC
C	DNA:RNA	≥ 50	67°C; 1xSSC -or- 45°C; 1xSSC, 50% formamide	67°C; 0.3xSSC
D	DNA:RNA	<50	T_D*; 1xSSC	T_D*; 1xSSC
E	RNA:RNA	≥ 50	70°C; 1xSSC -or- 50°C; 1xSSC, 50% formamide	70°C; 0.3xSSC
F	RNA:RNA	<50	T_F*; 1xSSC	T_F*; 1xSSC
G	DNA:DNA	≥ 50	65°C; 4xSSC -or- 42°C; 4xSSC, 50% formamide	65°C; 1xSSC
H	DNA:DNA	<50	T_H*; 4xSSC	T_H*; 4xSSC
I	DNA:RNA	≥ 50	67°C; 4xSSC -or- 45°C; 4xSSC, 50% formamide	67°C; 1xSSC
J	DNA:RNA	<50	T_J*; 4xSSC	T_J*; 4xSSC
K	RNA:RNA	≥ 50	70°C; 4xSSC -or- 50°C; 4xSSC, 50% formamide	67°C; 1xSSC
L	RNA:RNA	<50	T_L*; 2xSSC	T_L*; 2xSSC
M	DNA:DNA	≥ 50	50°C; 4xSSC -or- 40°C; 6xSSC, 50% formamide	50°C; 2xSSC
N	DNA:DNA	<50	T_N*; 6xSSC	T_N*; 6xSSC
O	DNA:RNA	≥ 50	55°C; 4xSSC -or- 42°C; 6xSSC, 50% formamide	55°C; 2xSSC
P	DNA:RNA	<50	T_P*; 6xSSC	T_P*; 6xSSC
Q	RNA:RNA	≥ 50	60°C; 4xSSC -or- 45°C; 6xSSC, 50% formamide	60°C; 2xSSC
R	RNA:RNA	<50	T_R*; 4xSSC	T_R*; 4xSSC

‡: The hybrid length is that anticipated for the hybridized region(s) of the hybridizing oligonucleotides. When hybridizing a oligonucleotide to a target oligonucleotide of unknown sequence, the hybrid length is assumed to be that of the hybridizing oligonucleotide. When oligonucleotides of known sequence are hybridized, the hybrid length can be determined by aligning the sequences of the oligonucleotides and identifying the region or regions of optimal sequence complementarity.

†: SSPE (1xSSPE is 0.15M NaCl, 10mM NaH_2PO_4, and 1.25mM EDTA, pH 7.4) can be substituted for SSC (1xSSC is 0.15M NaCl and 15mM sodium citrate) in the hybridization and wash buffers; washes are performed for 15 minutes after hybridization is complete.

*T_B - T_R: The hybridization temperature for hybrids anticipated to be less than 50 base pairs in length should be 5-10°C less than the melting temperature (T_m) of the hybrid, where T_m is determined according to the following equations. For hybrids less than 18 base pairs in length, T_m(°C) = 2(# of A + T bases) + 4(# of G + C bases). For hybrids between 18 and 49 base pairs in length, T_m(°C) = 81.5 + 16.6($\log_{10}$[Na^+]) + 0.41(%G+C) - (600/N), where N is the number of bases in the hybrid, and [Na^+] is the concentration of sodium ions in the hybridization buffer ([Na^+] for 1xSSC = 0.165 M).

It is clear to one of ordinary skill in the art that the contacting and hybridization steps can be optimized using any suitable method of optimization that is established in the art of hybridization. These include, but are not limited to, techniques that increase the efficiency of annealing or hybridization from complex mixtures of oligonucleotides (*e.g.*, PERT; *Nucleic Acids Research 23*:2339-2340, 1995) or hybridization in different formats (*e.g.*, using an immobilized template or using microtiter plates; *Analytical Biochemistry 227*:201-209, 1995).

Any parent polynucleotide with sufficient sequence similarity to the scaffold can be used to generate the donor fragments of the present invention. As defined herein, "sufficient sequence similarity" means that the sequence of the oligonucleotide need not reflect the exact sequence of the scaffold. Conditions are chosen to allow such sequences (and those having low similarity or similar sequences interrupted with dissimilar sequences) to hybridize the scaffold, such that double-stranded chimeric polynucleotides are formed. For example, non-complementary bases or insertions or deletions can be interspersed in sequences.

Upon contacting donor and scaffold fragments with each other, at least one hybridized complex is formed. Where flaps (unhybridized termini), gaps (single-stranded regions) and/or nicks occur in the hybridized complex, they can be trimmed, filled and ligated. In a particular embodiment, immediately adjacent oligonucleotides are ligated to each other. The term "adjacently hybridized" is used herein to describe the relative positions of two scaffold fragments hybridized to the same donor fragment, or two donor fragments hybridized to the same scaffold fragment, at positions such that only single-stranded sequence is contained between the two fragments. The term "immediately adjacently hybridized" is used herein to

describe adjacently hybridized scaffold or donor fragments that abut each other, *e.g.*, no intervening single-stranded sequence is contained between the two hybridized fragments.

Typically, a trimming, polymerization, ligation (TPL) step follows the contacting and hybridization of the population of donor fragments to the scaffold fragments. The TPL step includes trimming flaps, polymerization to fill in gaps between adjacently hybridized fragments, and ligation to join immediately adjacently hybridized fragments.

The utility of trimming flaps is realized because, in certain cases, the population of donor and scaffold fragments can hybridize such that at least one terminus of at least one of the hybridized fragments is unhybridized. The term "flaps" is used herein to describe the unhybridized terminus of an otherwise hybridized fragment. Internal sequences can also remain unhybridized, thus forming "loops" (loops are observed, for example, during denaturation/renaturation experiments with cDNA and genomic genes in which genomic introns loop out since there is no corresponding cDNA sequence to which to hybridize). The "trimming" of flaps, used herein to refer to a process of removing just the flaps, leaving the hybridized portion of the fragment intact, can be incorporated into the method of the present invention. Flaps can be trimmed enzymatically, *e.g.*, utilizing polymerases with single-stranded exonuclease activity or other single-stranded endonucleases or exonucleases, or chemically. The step of trimming flaps can be performed prior to or concurrently with the additional steps of polymerization and ligation.

Depending on specific hybridization capabilities, fragments can hybridize such that segments of the fragments remain unhybridized, *i.e.*, "gaps" are created. Such gaps could prevent the final formation of template-length chimeric polynucleotide, so a polymerization step is used to fill in the gaps. In a particular embodiment, the termini of the fragments are hybridized and at least one internal segment of a hybridized fragment is not hybridized.

Polymerization can be achieved either chemically or enzymatically. For example, gaps between adjacently hybridized fragments can be filled using a

suitable nucleic acid polymerizing enzyme, *e.g.*, a "polymerase". Suitable polymerases are commercially available. In one embodiment, gaps are filled in using prokaryotic, eukaryotic or viral polymerases. The polymerase can be thermostable or not thermostable. The polymerases can optionally have proof reading ability. Suitable polymerizing enzymes include T4 DNA polymerase, *Taq* DNA polymerase, *Pfu* DNA polymerase, Pol I, Klenow and Klenow 3'→5'^{exominus} (New England BioLabs, Beverly, MA). Typically, polymerases require a "primer" oligonucleotide that is extended by the polymerase in a process known as "strand extension." The polymerase reads the bottom strand and extends the primed top strand. A primer can be, for example, a short oligonucleotide that hybridizes to the bottom strand.

Control of enzymatic polymerization can be achieved, for example, by affecting the polymerase, *e.g.*, using a polymerase with altered processivity, or by affecting the template which is used by the polymerase during polymerization. For the purposes of the present invention, the gaps can be filled with or without the introduction of "errors" in comparison to the hybridized fragments.

In the method of the present invention, gaps between adjacently hybridized fragments can be separated by about 1,000 to about 100,000 template nucleotides. In other embodiments, the adjacently hybridized fragments are separated by about 500 to about 10,000 nucleotides; less than 1,000 nucleotides; less than 250 nucleotides; less than 50 nucleotides; or are separated by less than 25 nucleotides.

In another embodiment, gaps are filled in *in vivo,* wherein complexes containing oligonucleotides hybridized fragments are inserted or transformed into a suitable host cell. Gapped duplexes are examples of "self-priming" substrates for polymerases in the instances where the top strand contains an extendable 3' end and the single-stranded gap is used as the bottom strand that is read by the polymerase to extend the self-primed top strand.

In the method of the present invention, hybridized fragments are ligated. The hybridized fragments to be ligated are hybridized immediately adjacent to each other. The hybridized fragments are ligated using a suitable ligase. In one embodiment, ligation is accomplished using one or more ligases. Suitable ligases

include thermostable and non-thermostable ligases and include, but are not limited to, T4 DNA ligase, DNA ligase I, *Taq* ligase and *Tth* ligase. In another embodiment, ligation is accomplished using chemical means.

The final chimeric product can be a double-stranded chimeric polynucleotide that does not contain a contiguous, full-length template. It is therefore unnecessary to modify the template strand to facilitate its removal. This heteroduplex can be amplified using standard amplification techniques to generate homoduplex chimera or can be cloned and introduced into an organism using standard cloning and transformation techniques upon which replication *in vivo* will generate homoduplex chimeric molecules.

Chimeric polynucleotides can be selected or screened based on alterations of specific properties, *e.g.*, nucleotide structure, nucleotide function, altered enzymatic activities of proteins encoded by the chimeric polynucleotide, altered structural functions of proteins encoded by the chimeric polynucleotide, altered regulatory functions of proteins encoded by the chimeric polynucleotide, *etc.*, or a combination thereof. Subsequent selection and amplification of chimeric polynucleotides allows for the *in vitro* or *in vivo* directed evolution of biological molecules such as nucleic acid or polypeptides. This method for directed evolution would aid in the improvement of such molecules for use, for example, in medical therapies, as reagents in molecular biology, and in industry.

The present invention is particularly useful for evolving industrially or medically useful molecules for biochemical pathways, wherein the chimeric polynucleotide is itself a useful molecule (*e.g.*, promoter, aptamer, catalyst, enhancer or other regulatory element) or wherein the chimeric polynucleotide encodes a useful gene product. The chimeric polynucleotides can be or encode molecules that are more active under desired conditions to have altered or enhanced specificity, mutagenicity or fidelity. For example, desired conditions include conditions to which the reference molecule, oligonucleotide, template, or polypeptide encoded therein is not typically exposed or otherwise extreme conditions. Extreme conditions could include high or low temperature, extreme

high or low pH, extreme ionic strength, extreme solvent conditions such as organic solvent conditions, or a combination of two or more of these conditions.

Examples of industrially or medically useful polypeptides or oligonucleotides are well known in the art. Medically useful molecules include "bioactive" molecules, used herein to include peptides; proteins; polysaccharides and other sugars; lipids; and nucleic acid sequences, such as genes, and antisense molecules. Nucleic acid encoding enzymes that produce, modify or degrade polysaccharides, other sugars or lipids can be used as the scaffold, oligonucleotides or reference polynucleotide. Specific examples of bioactive molecules include, but are not limited to, insulin, erythropoietin, interferons, colony stimulating factors such as granulocyte colony stimulating factor, growth hormones such as human growth hormone, Insulin-Like Growth Factors I and II, Angiopoietin I and II, LHRH analogs, LHRH antagonists, tissue plasminogen activator, somatostatin analog, Factor VIII, Factor IX, calcitonin, dornase alpha, polysaccharides, AG337, bone inducing protein, bone morphogenic protein, brain derived growth factor, gastrin 17 immunogen, interleukins such as IL-2, PEF superoxide, permeability increasing protein-21, platelet derived growth factor, stem cell factor, thyrotropin, EGF, Tie-2 ligands, and somatomedin A and C.

One of skill in the art can readily select or design a scaffold to encode the molecule of interest to be evolved according to the method of the present invention. Methods for measuring activity of hormones, interleukins, growth factors and angiogenesis inhibitors and the like under desired conditions are well known in the art. One of ordinary skill in the art can readily determine the activity of the hormone, interleukin, growth factor or angiogenesis inhibitor encoded by the chimeric polynucleotide produced by the present invention and select those having the desired characteristics. Examples of medically useful molecules to be evolved according to the present invention also include enzymes that synthesize drugs, antibiotics, vitamins or co-factors. Other examples include vectors and genes for gene therapy. In addition, molecules that have desired therapeutic effect can be altered to lessen toxicity, antigenicity or other side effects.

Methods for determining activity under desired conditions include standard methods well known in the art. One of ordinary skill in the art can readily determine the activity of an enzyme encoded by a chimeric polynucleotide and select those oligonucleotides that encode enzymes that have desired characteristics. Enzymes include but are not limited to fermenting enzymes, proteases, lipases, oxidoreductases such as alcohol dehydrogenase, polymerases, hydrolases and luciferase.

Examples of industrially useful molecules include enzymes that synthesize polyketides, transform small molecules, hydrolyze substrates, replace steps in organic synthesis reactions or degrade pollutants such as aromatic hydrocarbons (*e.g.*, benzene, xylene, toluene and naphthalene), polychlorinated biphenyls and residual herbicides and pesticides. Catabolic pathways can be evolved using the present invention such that enzyme pathways are produced that degrade manmade pollutants that otherwise are not or only slowly catabolized. Oligonucleotides encoding such enzymes or fragments of coding regions can be used in the present invention as either the template, parent polynucleotides, a reference molecule to which chimeric polynucleotide products are compared, or combinations thereof. The method of the present invention can be used to increase, for example, the rate of an enzyme activity and the extent of the activity, *e.g.*, the affinity of the enzyme for its substrate. For example, the first enzyme in the metabolism of sulfur heterocycles by *Rhodococcus*, dibenzothiophene-monooxygenase (DBT-MO), is the bottleneck for both the rate and extent of sulfur oxidation in the biodesulfurization (BDS) process.

In one embodiment of the present invention, a chimeric polynucleotide is generated wherein one or more characteristics of the product molecule is different with respect to at least one reference polynucleotide. The difference in the chimeric polynucleotide can include a nucleotide change and/or amino acid changes in the encoded polypeptide in comparison to the reference polynucleotide, polypeptide or fragment thereof. The reference polynucleotide, polypeptide or fragment thereof can be the template or fragment, or can be a molecule related to the template used for comparison. For example, where the template is a non-functional version of a

oligonucleotide of interest or polypeptide encoded therein, then a reference molecule can be used for comparison to chimeric polynucleotides generated. The reference molecule can be a family member of the gene or gene product of interest, such as a homologous gene, or fragment thereof. One of skill in the art can readily choose a reference molecule based on the templates and oligonucleotides of interest used to generate the chimeric polynucleotides.

The characteristics to be altered according to the present invention include, but are not limited to, structural motif, stability, half-life, enzymatic activity, enzyme specificity, binding affinity, binding specificity, toxicity, antigenicity, interaction with an organism or interaction with components of an organism of the oligonucleotide or the encoded polypeptide. A functional characteristic can be altered according to the present invention such that the activity of said functional characteristic is enhanced at a higher or lower temperature compared to a reference molecule. Furthermore, said functional activities can be enhanced in various physical or chemical environments as described above or can be enhanced under standard conditions. Methods for measuring, selecting and screening these characteristics are well known in the art.

Structural motifs for proteins include, for example, α-helices, beta-sheets, solvent exposed loops, leucine zippers, β-barrel scaffolds and the like. Structural motifs for oligonucleotides include, for example, quadraplexes, aDNA, bDNA, zDNA, triple helices, stem loops, hairpins, protein binding sites and the like. Examples of regions are provided above. Methods for determining these motifs are well known in the art. In one embodiment, alteration of the characteristic includes an enhancement of the characteristic. In another embodiment, alteration of the characteristic includes a reduction in the characteristic.

In one embodiment of the present invention, a chimera is cloned prior to selection or screening. Methods of cloning oligonucleotides are well known in the art. Alternatively, the chimera can be selected or screened *in vitro* or *in vivo* prior to cloning.

The present invention allows the generation of at least one chimeric polynucleotide. The chimeric polynucleotides are different from any single

template used to generate the chimeric polynucleotide. Based on the method of the present invention, the differences can include, for example, an additional region, wherein the region is not present in the template. The additional region can be derived from an existing source of oligonucleotides, or a modified form thereof or can be a partially or completely random sequence. The additional region or regions can be present at either terminus of the resultant chimeric polynucleotide or can be present within the chimeric polynucleotide. Thus, the chimeric polynucleotide of the present invention can be longer than the template. In another embodiment, the chimeric polynucleotide can include an altered version of a region that is present in the template. The region can be the same length as the region in the hybridization template or can be longer or shorter than the region in the hybridization template. Thus, the chimeric polynucleotide can be the same size, longer or shorter than the template.

The invention will be further described with reference to the following non-limiting examples. The teachings of all the patents, patent applications and all other publications and websites cited herein are incorporated by reference in their entirety.

EXAMPLE 1

Method for Optimized Directed Evolution of PCI/TCI Polynucleotides

Heteroduplex Oligonucleotide Shuffling

Potato and tomato carboxypeptidase inhibitors (PCI and TCI, respectively) are 72 % identical at the amino acid level. To create a library of hybrid molecules from these two parents, three top strand oligonucleotides were synthesized to capture each polymorphism for the genes (Figure 2). Design modifications were carried out as described previously. Positioning of each oligonucleotide was selected to maximize the length of the perfectly base-paired interaction at the ends of each oligonucleotide without sacrificing representation of parental polymorphisms. Since no gaps were present, no polymerization was necessary and the top strand oligonucleotides were joined by ligation. DNA sequencing of 11

clones revealed between 1 and 7 crossovers per gene, with an average of 3.7 (ideal number of crossovers = 6). While each of the internal polymorphisms was represented at least once, representation of polymorphisms in the four positions nearest the junctures between the oligonucleotides were severely biased. Polymorphisms matching only the template gene were observed in 11 of 11 clones for three of these for positions and in 3 of 11 in the fourth.

The directed evolution of the PCI/TCI family of genes can be improved using synthetic oligonucleotides by optimizing the representation of allele single nucleotide polymorphisms (SNPs), dinucleotide polymorphisms (DiPs) and trinucleotide polymorphisms (TriPs) as alternative vs. degenerate loci. The mature coding regions of PCI/TCI are each 117 bp long and differ by 26 nucleotides (a 78% difference in sequence identity at the DNA level).

The PCI gene was altered to match common *E. coli* codon preferences (29 mutational changes). The TCI gene was altered in synonymous as well as non-synonymous codons. This resulted in a gene which was modified such that it contained 84% sequence identity with the original PCI gene (19 mismatches).

Mimicking *in vitro* recombination using standard degenerate oligonucleotides for these genes requires a two-fold degeneracy at each of these 19 positions, *i.e.*, to match one or the other parent, resulting in 2^{19} = 524,288-fold degeneracy. A minimum library size of over 1.5 million clones is required to capture each permutation of the parental alleles with a 95% degree of confidence. This large number is required whether a single degenerate oligonucleotide is generated or whether 19 degenerate oligonucleotides containing these 19 positions is generated. Although this number is an improvement when compared to the 2^{26}= 67 million clones which are necessary when the parents are not manipulated, further significant reductions in the required numbers would greatly increase efficiency. Focusing on the protein level, there are 11 amino acid residue differences between the two proteins. The following method of designing oligonucleotides balances the benefits of utilizing degenerate codons, *e.g.*, reduction of library size and screening, with the convenience of using commercially available synthetic methods (see Figure 2):

1. Where manipulation of parental sequences has allowed alternative codons at one locus to differ by a single nucleotide polymorphism (SNP), the alternative nucleotides at that single position are included in a two-fold degenerate locus in all oligonucleotides covering that region of the gene. The overall degeneracy of any particular oligonucleotide will be determined by the number of such SNPs and the chose termini of the oligonucleotide. These degenerate oligonucleotides will compete with alternative degenerate oligonucleotides described next. These alternative competitive oligonucleotides have identical termini.

2. Where alternative codons at a locus must differ by DiPs and TriPs, separate oligonucleotides are synthesized, each of which contain one or more of the possible permutations of the various DiPs and TriPs in the region encompassed by that oligonucleotide. For such oligonucleotides, too, the overall degeneracy is determined solely by the number of SNPs in that oligonucleotide. Since separate alternative oligonucleotides with the various permutations of DiPs and TriPs are otherwise identical, they will compete with each other for the same binding site. The termini of these oligonucleotides are identical to the desired degenerate codon oligonucleotides described above.

3. The oligonucleotides are designed to anneal perfectly at both termini to templates by synthesizing them to end in stretches of sequence identity between the two parents of, typically, 12 or more bases.

4. Other regions of the template are likewise hybridized to similarly designed degenerate and alternative degenerate oligonucleotides. Designing oligonucleotides that bind to other regions to include 5' phosphates and to abut perfectly with the neighboring oligonucleotides obviates the need for gap filling and flap trimming such that only the use of ligation is necessary to complete the chimeric strand. The need for forward and anchor oligonucleotides is also obviated, and the generation of parent clones by read-through from an upstream oligonucleotide is rendered unlikely.

For the trust of three primer binding sites, degenerate primer Deg1a is 5-fold degenerate and so consists of a mixture of 2^5=32 different primers. Since the 32 variations in Deg1b also compete for the same site, a total of 64 primers compete for this site (site 1). Likewise, there are four permutations of the four-fold degenerate Deg2 primers for a total of 16 comprising for site 2. Four permutations of 2-fold degeneracy indicate 8 primers competing for site 3. The total number of permutations of all the primers at each of the three sizes is 64 x 16 x 8=8192. Thus, the complete permutational diversity inherent in all the parental alleles can be captured in a theoretical library of 8192 clones. For 95% confidence in obtaining all of these clones, the library size (and the number of library clones screened) must be about 25,000.

EXAMPLE 2

Directed Evolution of *EGF* Gene Using TSTRAPS

Introduction

The method presented can generate every possible polymorphic permutation without bias by a protocol that involves annealing, polymerization and ligation of homoduplexed degenerate oligonucleotides. In preparation for the directed evolution of variant growth factors for differential signaling and inhibition of cellular proliferation in malignant cells, this method was applied to shuffle the genes encoding mouse and human epidermal growth factor (EGF), and to the simultaneous shuffling of EGF polymorphisms from five mammalian species. The resulting libraries of chimeric polynucleotides contained an unprecedented density of genetic crossovers and were completely free from genetic linkage. The mouse/human chimeric library represents the first gene family shuffled library to capture every possible permutation of the parental polymorphisms.

Results

Design modifications to the wild-type mouse and human EGF genes facilitates shuffling. Genes encoding the mature mouse and human EGF proteins are 74.5% identical. Modifications to these genes were made in order to allow

synthesis of optimal oligonucleotides for PARtially Scaffolded (PARSed) DNA shuffling. The design modifications include an upstream *Eco*RI site that allows for cloning of a gene encoding EGF as a fusion protein with the leader sequence of certain prokaryotic or eukaryotic expression/secretion vectors. Stop codons followed by a *Bam*HI cloning site were engineered downstream of the reading frame (Figure 3).

The design of the mouse and human *EGF* genes further included making the genes as similar as possible. This strategy required changing eleven silent polymorphisms in the mouse sequence to match the corresponding nucleotides in the human sequence. Six non-synonymous codons were also altered to reduce the polymorphic differences between them from an average of 2.5 to an average of 1, without changing the encoded amino acid residues (Figures 3 and 4). The number of nucleotide polymorphisms was thus reduced from 39 to 19, and the number of possible permutations of these clones from 239 to 219 (*i.e.*, from 5.5 x 1011 to 5 x 105 possible clones). The above manipulations reduced the total number of nucleotide permutations by six orders of magnitude without losing any of the polymorphic diversity inherent in the parental proteins.

PARSed DNA Shuffling Experimental Design

For the mouse/human *EGF* shuffling, oligonucleotides were synthesized to span the entire top strand of the modified *EGF* gene (Figure 4). Each oligonucleotide was designed to incorporate degeneracies that correspond to the polymorphisms of the mouse and human genes. In addition, polymorphic codons differing by two or three nucleotides in top strand chimeric oligonucleotides TS2 and TS3 were synthesized in separate reactions and then mixed to further reduce the degeneracy of the corresponding oligonucleotides by two-fold and four-fold, respectively. This last modification reduced the overall number of permutations needed to explore all the diversity of the wild-type parents to 6.5 x 104. Gaps of five and one nucleotide were allowed following TS1 and TS2, respectively, and thus required gap filling by DNA polymerase before ligation. TS2 and TS3 also possessed 5' phosphate groups to allow ligation. The top strand oligonucleotides

were positioned for gap filling and ligation by short bottom strand "scaffold" oligonucleotides. The scaffold oligonucleotides, however, possessed no 5' phosphate groups, and thus can not be ligated. The experimental design for the five-gene family shuffling is shown in Figure 4.

Analysis Of Mouse/Human PARSed DNA Shuffled Libraries

Products from the mouse/human PARSed DNA shuffled library were cloned. A total of 1010 chimeric genes were produced in a single PARSed shuffling reaction and a sampling of over 2 x 106 of these were captured in the first cloned library. DNA sequence analysis of random clones revealed only highly chimeric genes (Figure 5A). In 8 sequenced genes, the observed crossover density was 1 crossover per 17.5 bases, with an average of 7.75 crossovers per gene. These 8 clones also contained all 32 out of the 32 possible parental polymorphisms. Negative controls in which no polymerase or ligase was added to the PARSed DNA shuffling reaction yielded no product or clones. The distribution of polymorphisms from each parent at each polymorphic position clustered around the theoretical peak value of 50% (Figure 6). There was essentially no linkage between closely spaced parental polymorphisms. As discussed above, there are 6.5 x 104 unique permutations of the 32 polymorphisms. Since the above analysis indicates relatively little bias in generation of permutations, the number of clones needed to screen to have 99.99% probability of having screened every possible permutation in the library can be calculated. That number was calculated using the formula N = [ln(1-P)]/[ln(1-p-1)], where N is the number of screened clones, P is the probability of having screened any particular polymorphic permutation, and p is the number of possible permutations. Thus, screening 5.9 x 105 randomly chosen clones is required to screen, essentially to completion, every permutation of each parental polymorphism in these genes.

Analysis Of PARSed DNA Shuffled Libraries Of Five Mammalian Genes

EGF genes from human, mouse, rat, pig and horse differ in amino acid sequence identity by 58% to 84%. Top strand oligonucleotides were synthesized to

incorporate the polymorphisms of the parental genes and included design modifications as described above. Sequencing of 22 random clones from the chimeric library revealed crossovers between each of the 24 polymorphic positions. Seventeen of these clones are shown in Figure 5B. Single nucleotide deletions were observed in the other five clones and appear to represent artifacts within the synthesized oligonucleotide, TS5. Each of the 64 polymorphisms designed into the oligonucleotides were represented in this sampling. As was observed with the human/mouse shuffled EGF library, the frequency of crossovers between the closest alleles in these clones was the same as the frequency between the most distant alleles, and both classes centered around the ideal value of 50% (51% between the closest alleles and 50% between the most distal alleles). The number of crossovers per gene ranged from 6 to 18. The average number of crossovers in the library (11.0) differed from the theoretically perfect number of crossovers (23 crossover positions/2=11.5) by less than 5%.

Discussion

Optimal reassortment of polymorphisms in DNA shuffling is dependent on two factors. The first of these is crossover density. A typical pair of parental gene homologs that is 90% identical and only 1 kb in length will contain 100 polymorphic positions. Perfectly random recombination to explore all permutations of these polymorphisms would result in chimeric sequences averaging 50 crossovers per clone. Most other methods achieve an average of at most four crossovers for such genes. Moreover, generating multiple crossovers using current technologies becomes increasingly inefficient with decreasing gene size or increasing sequence divergence. Because of these limitations, the majority of classes of sequence permutations (*i.e.*, those involving more than a 1 crossover per 89 nucleotides (nt)) are left under-represented or entirely unexplored in the resulting chimeric libraries. The second critical parameter for optimizing recombination is the ability to achieve crossovers between close-lying polymorphisms (the ability to avoid genetic linkage effects). For hypothetical genes of 90% identity, the number of identical nucleotides between each polymorphism will average only nine bases. In the best

example reported to date, RACHITT generated 2.45 crossovers per gene between polymorphisms separated by 5 bp or fewer (Coco, W. *et al.*, *Nat. Biotechnol.* *19*:354-359, 2001). In contrast, PARSed DNA shuffling generated an average of 3.69 crossovers per gene between adjacent codons, and thus allows the testing of permutations of close-lying alleles that would otherwise tend to reassort as a single unit.

In PARSed DNA shuffling reactions, each ligation center involves three oligonucleotide participants- two top strands and a partial scaffold. Top strands that abut are ligated without polymerization. Strategically placed gaps are also used to reduce degeneracy of the annealed regions spanned by the partial scaffold. The degeneracies in the gap are introduced into the chimeric top strand during gap filling. Bottom strand oligonucleotides, *i.e.*, scaffold fragments, by contrast, are passive members in this particular embodiment of scaffolded shuffling. Bottom strand oligonucleotides can not be ligated to form a continuous strand because they do not contain, for example, a 5' phosphate. Alternatively, bottom strand oligonucleotides could be such that they can not be extended, *e.g.*, they could lack a 3' hydroxyl group. Bottom strand oligonucleotides, are not incorporated into the final library, and function only to guide homoduplex alignment of the top strands and as a source for sequence information in the small gapped regions.

The hybridizing regions of the bottom strand partial scaffolds did, in this example, contain degenerate positions. These degeneracies were designed to be perfectly complementary to the top strand chimeric oligonucleotide degeneracies at these positions. Because hybridization occurs during a gentle downward temperature ramp (*e.g.*, in this example, under conditions of high stringency), homoduplex annealing predominates over heteroduplex annealing. This encourages maximum binding strength even in regions of high sequence divergence and minimizes the required length of the scaffold, while simultaneously maximizing the specificity of binding and minimizing the representational bias of polymorphisms caused by mismatch discrimination.

Because polymorphisms are built into the degenerate oligonucleotide pools upon synthesis, physical crossovers between strands are not required. Shuffling of

the parental alleles results from ligation of any one oligonucleotide to a variety of alternative flanking oligonucleotides, as well as from polymerization across gaps in the degenerate partial scaffold oligonucleotides. Genetic linkage, a phenomenon that severely limits the sequence space explored by traditional shuffling methods, is thus absent in PARSed DNA shuffling. Recombination occurs between adjacent nucleotides as frequently as it does between distant polymorphisms. This feature allowed for the number of crossovers per gene to approach the ideal average and for the crossover density to reach 1 per 12 nt.

For the human/mouse EGF shuffling, libraries with a 1:1 ratio of the two alternative polymorphisms at each position were made. With random recombination, the libraries should have contained an ideal average number of crossovers equal to one-half of the number of potential crossover locations. This is five-fold higher than values reported for previous shuffling methods (Coco, W. *et al.*, *Nat. Biotechnol. 19*:354-359, 2001). The ideal average for the mouse/human EGF shuffling is thus 7.5 crossovers per gene. DNA sequence analysis of the PARSed DNA shuffling reaction revealed an essentially perfect average of 7.75 +/- 1.75 crossovers per gene. Similarly, the average observed for the 5-species DNA shuffled library was 11.0 +/- 2.2, which is statistically indistinguishable from the ideal number of 11.5. PARSed DNA shuffling is the first method to produce crossover densities as high as 1 per every 16 nt. It is also the first reported shuffling method that suffers no linkage effects, so that even higher crossover densities should be possible for more divergent parents. Every possible parental polymorphism in both the 2- and 5-species shuffled libraries was observed. In addition, the libraries approached the theoretical maximum of 50% reassortment at each polymorphism. These are the first gene-family DNA shuffled libraries to achieve this goal.

The unbiased linking of degenerate oligonucleotides is also important because it allows crossovers to approach the ideal distribution in short (*e.g.*, growth factor genes) or more divergent targets (as in our 5-gene library), where other multiple cross-over DNA shuffling methods become increasingly ineffective (Moore, G. *et al.*, *Proc. Natl. Acad. Sci. USA. 98*:3226-3231, 2001). To illustrate

this point, consider two close-lying alleles in two hypothetical gene homologs. Even if the chances of crossover are identical at each position along the entire length of the genes, the likelihood of crossovers between the two alleles is proportional to their separation and most unlikely for adjacent codons. In oligonucleotide based molecular breeding, the segregation of alleles can potentially be 50%, regardless of separation. This level of non-linkage, however, was not observed using oligonucleotide-based methods that rely on heteroduplex annealing, *e.g.*, RACHITT™. In contrast to the heteroduplex annealing process, the present homoduplex method allowed representation of all alleles at a frequency centered near the theoretically perfect 50%. Additionally, since each of the starting oligonucleotides contained polymorphisms from multiple parents, there was no chance of getting a significant proportion of unshuffled parental clones in the shuffled library.

Figure 4C, depicts an oligonucleotide shuffling format involving annealing of degenerate oligonucleotides to a gene-length transient template. Complex chimeric libraries can be generated in this way. A requirement for heteroduplex annealing in such methods, however limits the utility of this approach for the divergent genes used in family shuffling. Heteroduplex hybridization in divergent regions involves a compromise between polymorphism bias through mismatch discrimination under stringent annealing conditions on the one hand, and an increased proportion of non-specific products under less stringent conditions on the other. To avoid this bias, some polymorphisms must be eliminated in order to generate perfectly hybridizing anchors or "sticky feet" at the ends of each oligonucleotide. Similarly, the limitations of family shuffling by sexual PCR and other methods are well characterized. These can include generation of non-specific products, retention of unshuffled clones in the final chimeric library, severe linkage effects and, with one exception (Coco, W. *et al.*, *Nat. Biotechnol. 19*:354-359, 2001), limitation to four or fewer crossovers per gene.

Unlike other shuffling methods, PARSed DNA shuffling involves no thermocycling, stuttering, heteroduplex annealing or unmodified parental gene fragments. The single event, high stringency homoduplex hybridization shuffling

method will result in a diverse, unbiased chimeric gene library. The properties of PARSed DNA shuffling circumvent or minimize limitations of other mutagenesis or shuffling methods that rely on heteroduplex formation for gene family shuffling. The generated libraries described herein contained no observed bias, linkage or unwanted sibling and parental clones. The total number of possible permutations of the mouse/human EGF polymorphisms is 6.5 x 104. To capture 99.99% of these permutations in a random, unbiased library would require 5.9 x 105 members. Therefore, 2 x 106 chimeric *EGF* genes were cloned for this library. This is the first example of DNA shuffling that has been demonstrated to fully capture every possible parental permutation in a chimeric gene family library. The utility of design modifications that allowed for facilitated shuffling is not restricted to the examples presented here. Rather, they should be broadly applicable to any polynucleotide of interest or shuffling method. While the current application involved shuffling of small growth factor genes, it is amenable to larger sequences. Oligonucleotide-based gene synthesis protocols have been used for genes that are >1.5 kb. PARSed DNA shuffling is directly adaptable to such sizes, however for larger sequences it may be necessary to shuffle subsets of the genes that can subsequently be linked to give a full length product.

The goal of DNA shuffling is to create libraries of molecules that explore some random subset of all of the sequence space that is generated by the permutations of polymorphisms from two or more parental polynucleotides. This enormous variety of possible permutations provides a vast, diverse pool of functional protein variants from which improved protein characteristics can be selected or screened. Eliminating bias in the reassortment of polymorphisms is necessary to achieve the broadest and most representative search of the genetic diversity inherent in parental polynucleotides. The use of homoduplexed degenerate oligonucleotides to shuffle polynucleotides has achieved this goal for the genes presented herein and should be applicable to a broad range of nucleotide and protein engineering problems.

Experimental Protocols

Degenerate/alternate synthetic oligonucleotides. TS1-3 and partial scaffold (PS) 1 and 2 oligonucleotides were synthesized. Degenerate positions are indicated using IUPAC abbreviations. Otherwise identical oligonucleotides with alternative codons are distinguished by letter suffixes A and B. Oligonucleotides were synthesized by Sigma-Genosys(The Woodlands, Texas).

(SEQ ID 43) TS1A: 5'OH-
gcgcaggccggaattcagaatagtKatYctgRatgtccctYgtccYatgatgggtactgcctc

(SEQ ID 44) TS2A: 5'PO_4-
tggtgtgtgcatgYatattgaaKcattggacaagtatRcatgcaactgtgttRttggctaca

(SEQ ID 45) TS2B: 5'PO_4-
tggtgtgtgcatgYatattgaaKcattggacagctatRcatgcaactgtgttRttggctaca

(SEQ ID 46) TS3A: 5'PO_4-
cggggaKcgatgtcagtaccgagacctgaRgtggtgggaactgcgctaataggatccggctga
gcaccgcgc

(SEQ ID 47) TS3B: 5'PO_4-
cgcggggaKcgatgtcagactcgagacctgaRgtggtgggaactgcgctaataggatccggct
gagcaccgcgc

(SEQ ID 48) PS1: 5'OH- ctgacatcgMtccccgMtgtagccaaYaacacagttgcatg

(SEQ ID 49) PS2: 5'OH- ttcaatatRcatgcacacaccaYcatKgaggcagtacccatcat

Each "B" alternate oligonucleotide was combined with its "A" counterpart in equimolar amounts. The resulting five populations (TS1-3 and PS1/2) were then combined in equimolar amounts and diluted to 0.625 mM in annealing buffer.

PARSed DNA shuffling using thermophilic enzymes

Annealing was performed in 1X *Thermus aquaticus* (*Taq*) ligase buffer (NEB) supplemented with 2 mM dNTPs. The temperature was brought to 84°C for 1 minute, cooled rapidly to 75°C, ramped to 45°C over 50 minutes, and then brought rapidly to 65°C. *Taq* DNA ligase (40 U) and 0.5 U *Taq* DNA polymerase

were then added and incubated at 65°C for 40 minutes. The reaction was stopped by freezing. The resulting chimeric top strands were amplified by PCR and cloned. As a control, polymerase and ligase were omitted during the oligonucleotide assembly reactions. Subsequent PCR yielded a mixture of low molecular weight, non-specific DNA fragments. No full-length *EGF* genes were detectable upon cloning of these products.

The teachings of all references, patents and patent applications cited herein are hereby incorporated by reference in their entireties. While this invention has been particularly shown and described with references to preferred embodiments thereof, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the scope of the invention encompassed by the appended claims.

CLAIMS

What is claimed is:

1. A method for forming a chimeric polynucleotide comprising:

contacting a population of single-stranded scaffold fragments with a population of donor fragments under conditions such that at least one scaffold fragment hybridizes to at least two donor fragments at distal regions of the scaffold fragment;

treating the hybridized complexes such that single-stranded regions of the hybridized complex are filled-in; and

treating the filled-in hybridized complexes such that adjacent fragments are ligated, forming a chimeric polynucleotide.

2. The method of Claim 1, further comprising the step of trimming flaps prior to ligation.

3. The method of Claim 1, wherein the scaffold fragments comprise sequences of from about 10 to about 1000 nucleotides in length.

4. The method of Claim 1, wherein the population of scaffold fragments is derived from a single strand of a parent polynucleotide.

5. The method of Claim 1, wherein the donor fragments comprise sequences of about 10 to about 1000 nucleotides in length.

6. The method of Claim 1, wherein the donor fragments are single-stranded.

7. The method of Claim 6, wherein the population of donor fragments is derived from a single strand of a parent polynucleotide.

8. The method of Claim 1, wherein the at least one scaffold and the at least two donor fragments hybridize to each other under conditions of low stringency.

9. The method of Claim 1, wherein the population of scaffold fragments and the population of donor fragments are produced synthetically.

10. The method of Claim 1, wherein the population of scaffold fragments and the population of donor fragments are produced by cleaving a polynucleotide of interest that is a full length cDNA.

11. The method of Claim 1, wherein at least one of the fragments of the scaffold or donor populations comprises at least one region of random sequence.

12. The method of Claim 1, further comprising a step of preparing at least one single-stranded population of scaffold fragments, derived from a randomly fragmented single-stranded polynucleotide of interest.

13. The method of Claim 1, wherein the populations of scaffold and donor fragments are sufficient to form a full-length chimeric polynucleotide.

14. The method of Claim 1, further comprising screening or selecting at least one chimeric polynucleotide having desired characteristics.

15. A chimeric polynucleotide prepared according to the method of Claim 1.

16. A library of chimeric polynucleotides prepared according to the method of Claim 1.

17. The library of Claim 16, wherein the majority of the chimeric polynucleotides contain at least 3 crossover sites.

18. The library of Claim 17, wherein at least one chimeric polynucleotide contains the number of crossovers within 10% of the theoretical limit.

19. The library of Claim 18, wherein at least five chimeric polynucleotides contain the number of crossovers within 10% of the theoretical limit.

20. A method for forming at least one double-stranded chimeric polynucleotide having desired characteristics comprising:

contacting a population of scaffold fragments derived from a template polynucleotide with a population of donor fragments under conditions such that fragments of the scaffold and donor populations can hybridize to each other;

forming at least one hybridized complex comprising at least one scaffold fragment hybridized to at least two donor fragments;

treating the hybridized complex such that single-stranded regions of the hybridized complex are filled-in;

treating the filled-in hybridized complex such that adjacent fragments are ligated,

thereby forming a double-stranded chimeric polynucleotide.

21. The method of directed evolution, comprising screening or selecting at least one double-stranded chimeric polynucleotide from the library of Claim 20 having desired characteristics.

22. The method of Claim 20, further comprising trimming flaps.

23. The method of Claim 20, wherein the scaffold fragments comprise sequences that are a maximum of 25 percent as long as a polynucleotide of interest.

24. The method of Claim 20, wherein the scaffold fragments comprise sequences of from about 25 to about 1000 nucleotides in length.

25. The method of Claim 20, wherein the donor fragments comprise sequences of from about 25 to about 1000 nucleotides in length.

26. The method of Claim 20, wherein the donor fragments are single-stranded.

27. The method of Claim 26, wherein the population of donor fragments is derived from a single strand of a parent polynucleotide.

28. The method of Claim 20, wherein the scaffold and donor fragments hybridize to each other under conditions of low stringency.

29. The method of Claim 20, wherein the single-stranded regions are filled in using a polymerase.

30. The method of Claim 20, wherein the hybridized fragments are ligated using *Taq* DNA ligase or T4 DNA ligase.

31. The method of Claim 20, further comprising repeating steps hybridizing, filling in and ligating, wherein one or more chimeric polynucleotides is used to generate the populations of scaffold or donor fragments.

32. The method of Claim 20, wherein at least one of the fragments of the scaffold or donor populations comprises at least one region of random sequence.

33. A chimeric polynucleotide prepared according to the method of Claim 20.

34. A method for preparing a population of scaffold fragments, comprising the steps of:

amplifying a oligonucleotide of interest in a polymerase chain reaction, wherein the 5' terminus of a first primer comprises a 5' phosphate and wherein the 5' terminus of a second primer is devoid of a 5' phosphate;

contacting the amplified oligonucleotide with lambda exonuclease under conditions wherein oligonucleotides having a 5' phosphate are digested, leaving single-stranded oligonucleotides; and

fragmenting the single-stranded oligonucleotides, thereby preparing a population of scaffold fragments.

35. A method for forming a chimeric polynucleotide comprising:

treating a library of oligonucleotide fragments derived from a parent polynucleotide of interest and allelic variations thereof, wherein the population of fragments comprises a first population of oligonucleotides derived from one strand of the parent polynucleotide and allelic variations thereof and oligonucleotides of a second population wherein oligonucleotides are synthesized *in vitro* and derived from the other strand of the known parent polynucleotide and allelic variations thereof under conditions such that oligonucleotides of the first population can hybridize to oligonucleotides of the second population to form a gapped homoduplex;

treating the gapped homoduplex with a polymerase, wherein polynucleotide strand extension produces a double-stranded polynucleotide comprising at least one nicked strand; and

treating the nicked polynucleotide with a ligase,

thus forming a full-length polynucleotide.

36. A method of forming a single-stranded chimeric polynucleotide according to the method of Claim 35, wherein the oligonucleotides of the second population do not contain a 5' phosphate group, further comprising the step of removing the oligonucleotides of the second population after ligation.

37. The method of Claim 32, comprising the additional step of amplifying the single-stranded chimeric polynucleotide in a nucleic acid amplification reaction thereby producing more than one copy of a double-stranded chimeric polynucleotide.

38. A method of forming a single-stranded chimeric polynucleotide according to the method of Claim 35, wherein the oligonucleotides of the second population do not contain a 3' hydroxyl group, further comprising the step of removing the oligonucleotides of the second population after ligation.

39. The method of Claim 37, comprising the additional step of amplifying the single-stranded chimeric polynucleotide in a nucleic acid amplification reaction thereby producing more than one copy of a double-stranded chimeric polynucleotide.

40. The method of Claim 39, wherein the gapped homoduplex is full-length.

41. The method of Claim 35, wherein the known parent molecule sequence is from about 50 bases to about 2 kilobases in length.

42. The method of Claim 35, wherein the known parent sequence is from about 1 kilobase to about 5 kilobases in length.

43. The method of Claim 35, wherein the known parent sequence is from about 2 kilobases to about 25 kilobases in length.

44. The method of Claim 35, comprising an additional recombination step between the chimeric polynucleotide and a parent molecule or allelic variation thereof.

45. A library of chimeric polynucleotides comprising more than one chimeric polynucleotides formed according to the method of Claim 35.

46. The method of Claim 35, wherein the oligonucleotides of the second population are derived from regions of sequence identity between parent polynucleotides and allelic variations thereof.

47. The method of Claim 35, wherein the gapped homoduplex contains polymorphic sites in at least one double-stranded region of the homoduplex.

48. The method of Claim 35, wherein the gapped homoduplex contains at least one polymorphic site in the gapped region of the gapped homoduplex.

49. A method for directed evolution comprising:

forming a library of chimeric polynucleotides comprising:

contacting a first population of oligonucleotides with a second population of oligonucleotides, wherein the sequences of the first and second oligonucleotide populations are complementary to one another, under conditions such that oligonucleotides of the first population can hybridize to oligonucleotides of the second population to form a gapped homoduplex;

treating the gapped homoduplex with a polymerase, wherein polynucleotide strand extension produces a nicked polynucleotide;

treating the nicked polynucleotide with a ligase, such that nicks are ligated; and

screening the library of chimeric polynucleotides for a characteristic of interest.

50. The method of Claim 49, wherein the oligonucleotides of the first population and the oligonucleotides of the second population are derived from a known polynucleotide of interest.

51. The method of Claim 50, further comprising repeating the steps using the chimeric polynucleotide as the known polynucleotide of interest in the subsequent round of directed evolution.

52. The method of Claim 51, wherein the steps are repeated from about 2 to 50 times using a screened population of chimeric polynucleotides as the parent polynucleotides used to generate scaffold and donor fragments in a subsequent round of directed evolution.

53. The method of Claim 49, wherein the oligonucleotides of the second population do not contain 5' phosphate groups.

54. The method of Claim 49, wherein the oligonucleotides of the second population do not contain 3' hydroxyl groups.

55. The method of Claim 49, wherein the screening step comprises screening the function of the transcribed and/or translated products of the library of chimeric polynucleotides.

56. The method of Claim 49, comprising cloning the library of chimeric polynucleotides into a suitable vector prior to the screening step.

57. The method of Claim 49, further comprising:

cloning the chimeric polynucleotides into expression vectors;

transforming a suitable cell line with the cloned chimeric polynucleotides;

inducing expression of the cloned chimeric polynucleotide;

assaying the expressed product for a characteristic of interest; and

selecting the chimeric polynucleotide that expressed products with an improved characteristic of interest.

58. The method of Claim 49, further comprising:

transcribing and translating the chimeric polynucleotide *in vitro*;

assaying the transcribed and translated products for a characteristic of interest; and

selecting the chimeric polynucleotide that lead to transcribed and translated products with an improved characteristic of interest.

59. A chimeric polynucleotide formed and selected according to the method of Claim 49.

60. A method for forming a single-stranded chimeric polynucleotide comprising:

treating a library of oligonucleotide fragments derived from a parent polynucleotide of interest and allelic variations thereof, wherein the population of fragments comprises a first population of oligonucleotides derived from one strand of the parent polynucleotide and allelic variations thereof and oligonucleotides of a second population wherein oligonucleotides are synthesized *in vitro* and derived from the other strand of the known parent polynucleotide and allelic variations thereof under conditions and wherein oligonucleotides of the second population do not contain 5' phosphate groups such that oligonucleotides of the first population

can hybridize to oligonucleotides of the second population to form a gapped homoduplex;

treating the gapped homoduplex with a polymerase, wherein polynucleotide strand extension produces a double-stranded polynucleotide comprising at least one nicked strand;

treating the nicked polynucleotide with a ligase, such that the first population of oligonucleotides are ligated and the second population of oligonucleotides are not ligated; and

removing the hybridized oligonucleotides of the second population,

thus forming a single-stranded chimeric polynucleotide.

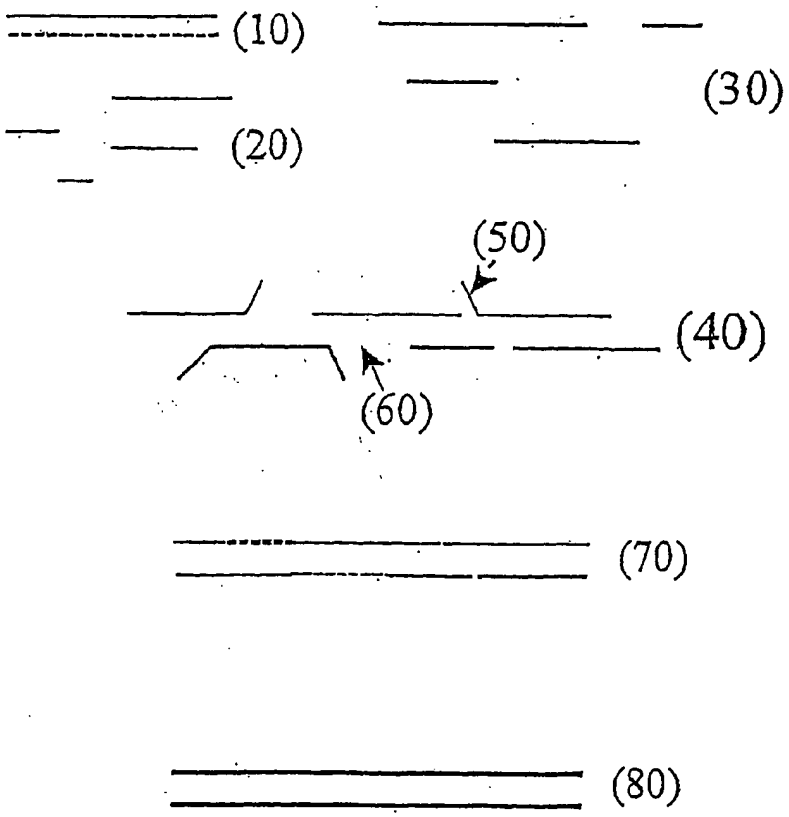


Fig. 1

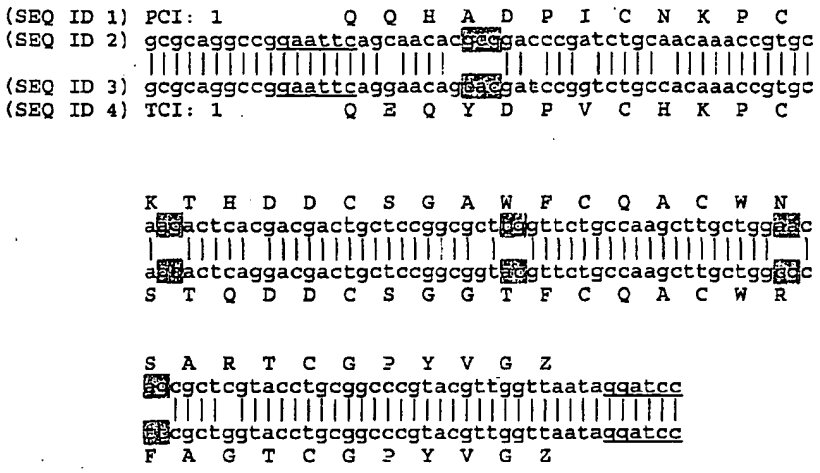

(SEQ ID 1) PCI: 1 Q Q H A D P I C N K P C
(SEQ ID 2) gcgcaggccggaattcagcaacac ggacccgatctgcaacaaaccgtgc
(SEQ ID 3) gcgcaggccggaattcaggaacag gatccggtctgccacaaaccgtgc
(SEQ ID 4) TCI: 1 Q Z Q Y D P V C H K P C
K T H D D C S G A W F C Q A C W N
S T Q D D C S G G T F C Q A C W R
S A R T C G Z Y V G Z
F A G T C G Z Y V G Z

Fig. 2A

Deg1a:
5'-phosphate -gcgcaggccggaattcag (c/g) aaca (c/g) gcgga (c/t) ccg (a/g) tctgc (a/c) acaaac
46-mer (SEQ ID NO 5)

Deg1b:
5'-phosphate -gcgcaggccggaattcag (c/g) aaca (c/g) tacga (c/t) ccg (a/g) tctgc (a/c) acaaac
46-mer (SEQ ID NO 6)

Deg2a:
5'-phosphate -cgtgcaagactca (c/g) gacgactgctccggcg (c/g) ttggttctgccaa
44-mer (SEQ ID NO 7)

Deg2b:
5'-phosphate -cgtgcaagactca (c/g) gacgactgctccggcg (c/g) tacgttctgccaa
44-mer (SEQ ID NO 8)

Deg2c:
5'-phosphate -cgtgcagcactca (c/g) gacgactgctccggcg (c/g) ttggttctgccaa
44-mer (SEQ ID NO 9)

Deg2d:
5'-phosphate -cgtgcagcactca (c/g) gacgactgctccggcg (c/g) ttacttctgccaa
44-mer (SEQ ID NO 10)

Deg3a:
5'-phosphate -gcttgctggaacagcgct (c/g) gtacctgcggcccgtacgttggttaata
47-mer (SEQ ID NO 11)

Deg3b:
5'-phosphate -gcttgctggaacttcgct (c/g) gtacctgcggcccgtacgttggttaata
47-mer (SEQ ID NO 12)

Deg3c:
5'-phosphate -gcttgctggcgcagcgct (c/g) gtacctgcggcccgtacgttggttaata
47-mer (SEQ ID NO 13)

Deg3d:
5'-phosphate -gcttgctggcgcttcgct (c/g) gtacctgcggcccgtacgttggttaata
47-mer (SEQ ID NO 14)

Fig. 2B

A)

```
                        EcoR1    >
human(SEQ ID 15)GCGCAGGCCGGAATTCAGAATAGTGACTCTGAATGTCCCCTGTCCCACGATGGGTACTGC
mouse(SEQ ID 16)-------------------------T-TC-A-G---C--ATCC--AT-T-----A------
human(SEQ ID 17)                    AsnSerAspSerGluCysProLeuSerHisAspGlyTyrCys
mouse(SEQ ID 18)                          TyrProGly       Ser    Tyr

human           CTCCATGATGGTGTGTGCATGTATATTGAAGCATTGGACAAGTATGCATGCAACTGTGTT
mouse           ---A---G---C---------C--------T--C------GC--CA--------------
human           LeuHisAspGlyValCysMetTyrIleGluAlaLeuAspLysTyrAlaCysAsnCysVal
mouse             AsnGly             Eis       Ser       Ser    Thr

human           GTTGGCTACATCGGGGAGCGATGTCAGTACCGAGACCTGAAGTGGTGGGAACTGCGCTAA
mouse           A-------TTCT-----T---------ACT--------ACGA--------G-----T---
human           ValGlyTyrIleGlyGluArgCysGlnTyrArgAspLeuLysTrpTrpGluLeuArgStp
mouse           Ile       Ser    Asp          Thr          Arg

                   BamH1
human           TAGGATCCGGCTGAGCACCGCGC
mouse           ------------------------
```

B)

```
                           EcoR1    >
human(SEQ ID 19)GCGCAGGCCGGAATTCAGAATAGTGATTCTGAATGTCCCTTGTCCCACGATGGGTACTGC
mouse(SEQ ID 20)-------------------------T--C---G--------C----T--------------
human(SEQ ID 21)                    AsnSerAspSerGluCysProLeuSerHisAspGlyTyrCys
mouse(SEQ ID 22)                          TyrProGly       Ser    Tyr

human           CTCCATGATGGTGTGTGCATGTATATTGAAGCATTGGACAAGTATGCATGCAACTGTGTT
mouse           ---A---G-------------C--------T---------GC---A--------------
human           LeuHisAspGlyValCysMetTyrIleGluAlaLeuAspLysTyrAlaCysAsnCysVal
mouse             AsnGly             His       Ser       Ser    Thr

human           GTTGGCTACATCGGGGAGCGATGTCAGTACCGAGACCTGAAGTGGTGGGAACTGCGCTAA
mouse           A---------C------,T---------ACT----------G-------------------
human           ValGlyTyrIleGlyGluArgCysGlnTyrArgAspLeuLysTrpTrpGluLeuArgStp
mouse           Ile       Ser    Asp          Thr          Arg

                   BamH1
human           TAGGATCCGGCTGAGCACCGCGC
mouse           ------------------------
```

Fig. 3

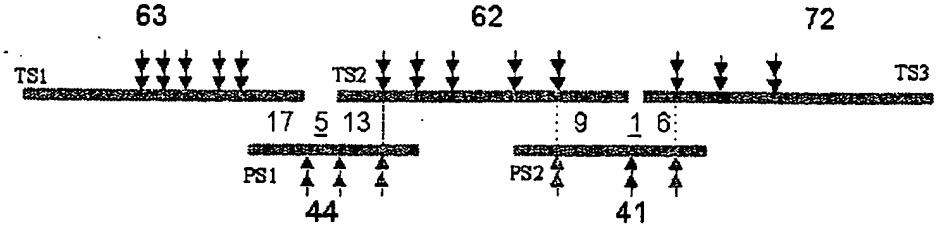


Fig. 4A

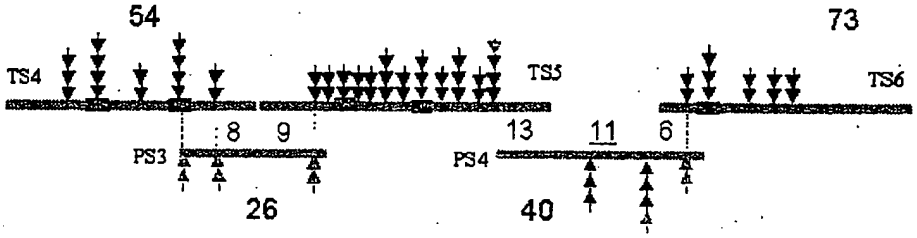


Fig. 4B

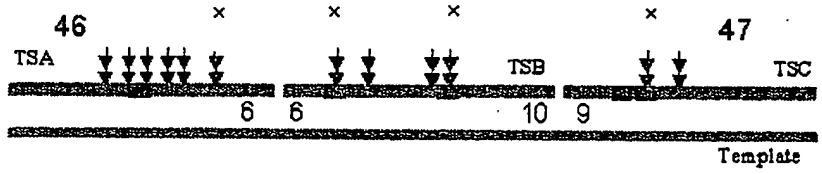


Fig. 4C

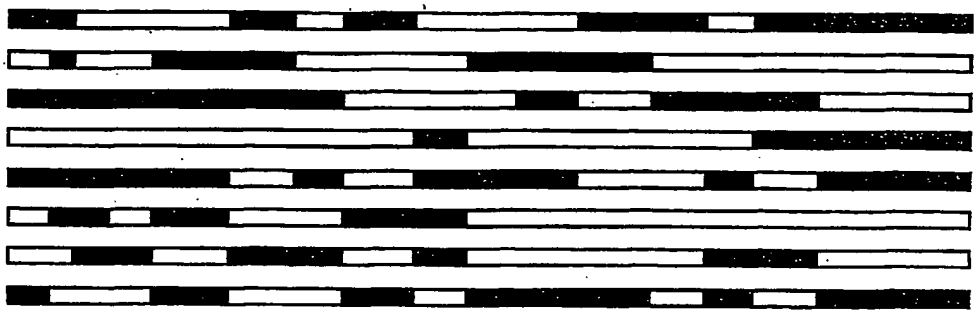


Fig. 5A

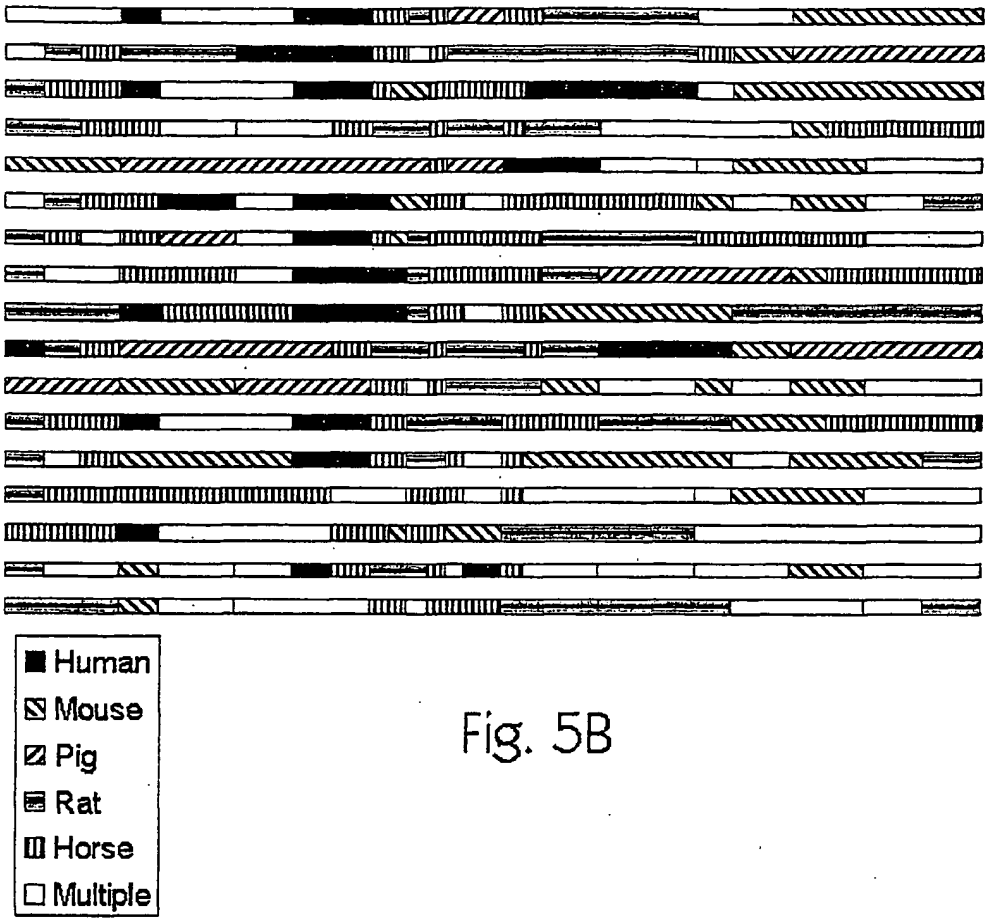


Fig. 5B

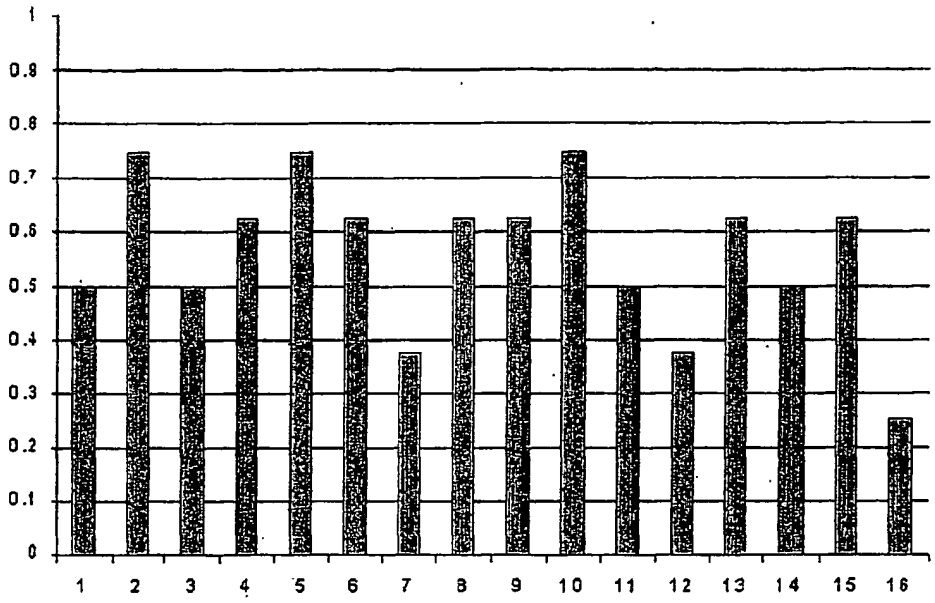


Fig. 6